全方位架构企业 赢在大数据运营

Big Data Operations

The New Thinking of Service-Oriented Enterprise Architecture

大数据运营

服务型企业架构新思维

李福东◎著

清華大學出版社

大数据运营

服务型企业架构新思维

李福东◎著

清華大學出版社
北 京

内 容 简 介

犹如个人的修齐治平，企业大数据运营同样需要经历筑巢、联姻、孕育、分娩、培育、腾飞 6 个阶段。筑巢的目的是建立一个结构严谨的企业架构，为企业发展打下基础。联姻是将企业架构与大数据结合起来，从业务活动角度提出对大数据的需求，从大数据角度提出对业务活动的支撑方法与过程。孕育是以大数据战略为驱动，构建大数据应用框架。分娩是将大数据从想象变为现实，形成可以运行的大数据服务。培育是根据新需求对大数据服务进行优化，更加有效地支撑企业业务活动。腾飞指的是大数据服务在行业中的应用，企业在大数据服务的辅助下走向成功和辉煌。

图书在版编目（CIP）数据

大数据运营：服务型企业架构新思维 / 李福东著. —北京：清华大学出版社，2015
ISBN 978-7-302-40537-5

Ⅰ. ①大…　Ⅱ. ①李…　Ⅲ. ①企业管理-信息管理-研究　Ⅳ. ①F272.7

中国版本图书馆 CIP 数据核字（2015）第 137543 号

责任编辑： 冯志强　薛　阳
封面设计： 吕单单
责任校对： 胡伟民
责任印制： 沈　露

出版发行： 清华大学出版社
网　　址： http://www.tup.com.cn, http://www.wqbook.com
地　　址： 北京清华大学学研大厦 A 座　　**邮　　编：** 100084
社 总 机： 010-62770175　　**邮　　购：** 010-62786544
投稿与读者服务： 010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈： 010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者： 三河市中晟雅豪印务有限公司
经　　销： 全国新华书店
开　　本： 185mm×230mm　　**印　　张：** 20.75　　**插　　页：** 3　　**字　　数：** 518 千字
版　　次： 2015 年 8 月第 1 版　　**印　　次：** 2015 年 8 月第 1 次印刷
印　　数： 1～4000
定　　价： 59.00 元

产品编号：061327-01



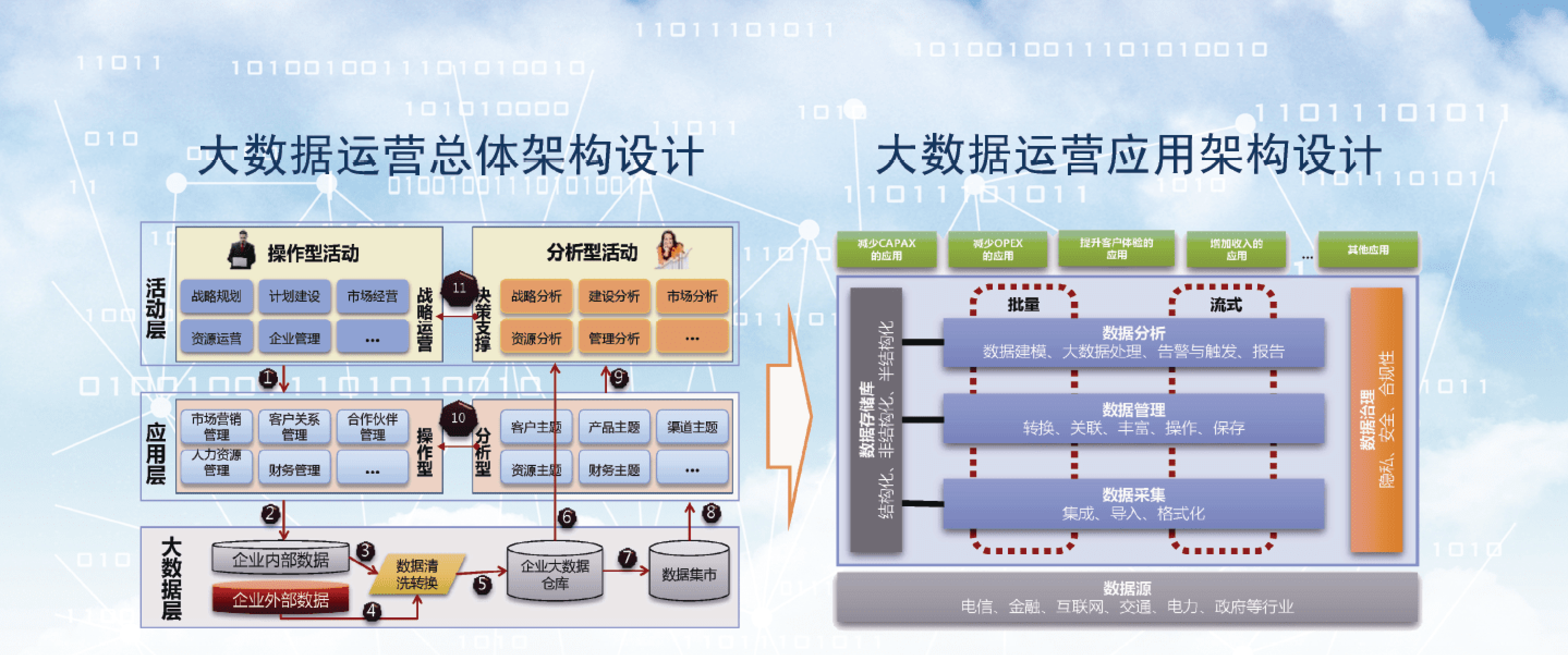
多视角的企业架构模型



企业架构模型10维架构的价值和作用

企业架构总体模型

- 管 10 治理架构：有规矩才能有方圆
- 安 9 安全架构：开放和共享的代价
- 实 8 部署架构：将软件与硬件结合
- 力 7 技术架构：创新的驱动力和源动力
- 模 6 数据架构：信息的模型化和规范化收敛
- 动 1 过程架构：时空交叉的矩阵框架
- 静 2 信息架构：业务活动的静态表述
- 桥 3 应用架构：连接业务与技术的纽带
- 合 4 集成架构：业务与技术的粘合剂
- 口 5 功能架构：应用能力在系统层面的落地



大数据服务从业务到系统的落地过程

业务层

业务活动 → 分解 → 决策活动（知）/ 执行活动（行） → 实现 → 知行合一

能力层

大数据服务需求 = 大数据能力分析 + 问题域特点分析

技术层

在线还是离线？

实时还是非实时？

？？？

批量还是流式？

关系型还是分布式？

大数据服务开发测试环境

小数据存储区
主机
磁盘阵列
磁盘阵列
开发区
应用服务器
数据库服务器
展示服务器
开发工程师
ETL/ELT区
2.2 辅助分析数据
2.1 样本大数据
6 数据颗粒度加粗
3.2
3.1
5
外网用户
1 数据装载
数据工程师
大数据存储区
目录节点
数据节点
数据节点
4.2
测试区
应用服务器
数据库服务器
展示服务器
4.1
测试工程师

行业大数据的汇聚成为大数据服务的构建基石

通信大数据

通话记录
上网记录
客户属性
号卡数据
……

金融大数据

交易记录
信用记录
个人属性
企业属性
……

互联网大数据

搜索记录
社交网络记录
购物记录
点评记录
……

行业外大数据

交通行业	车辆信息、驾驶员信息、维保记录、违规记录、交通事故记录、年检记录、…
⋮	
公共事业	用水记录、用电记录、煤气使用记录、纳税记录、报警记录、违法记录、…

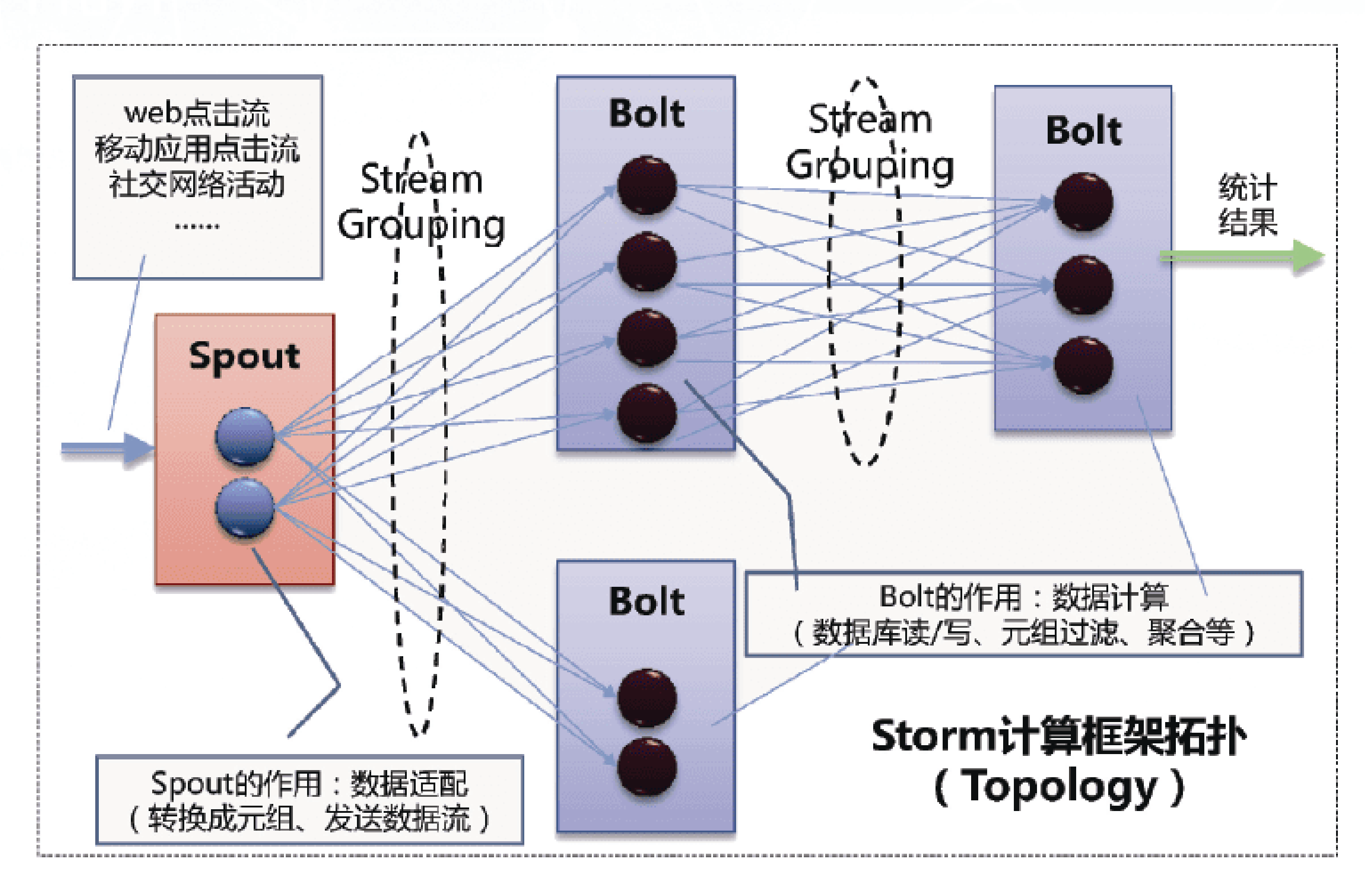
Frameworx总体架构

业务过程框架
集成框架
应用框架
信息框架
最佳实践
业务度量

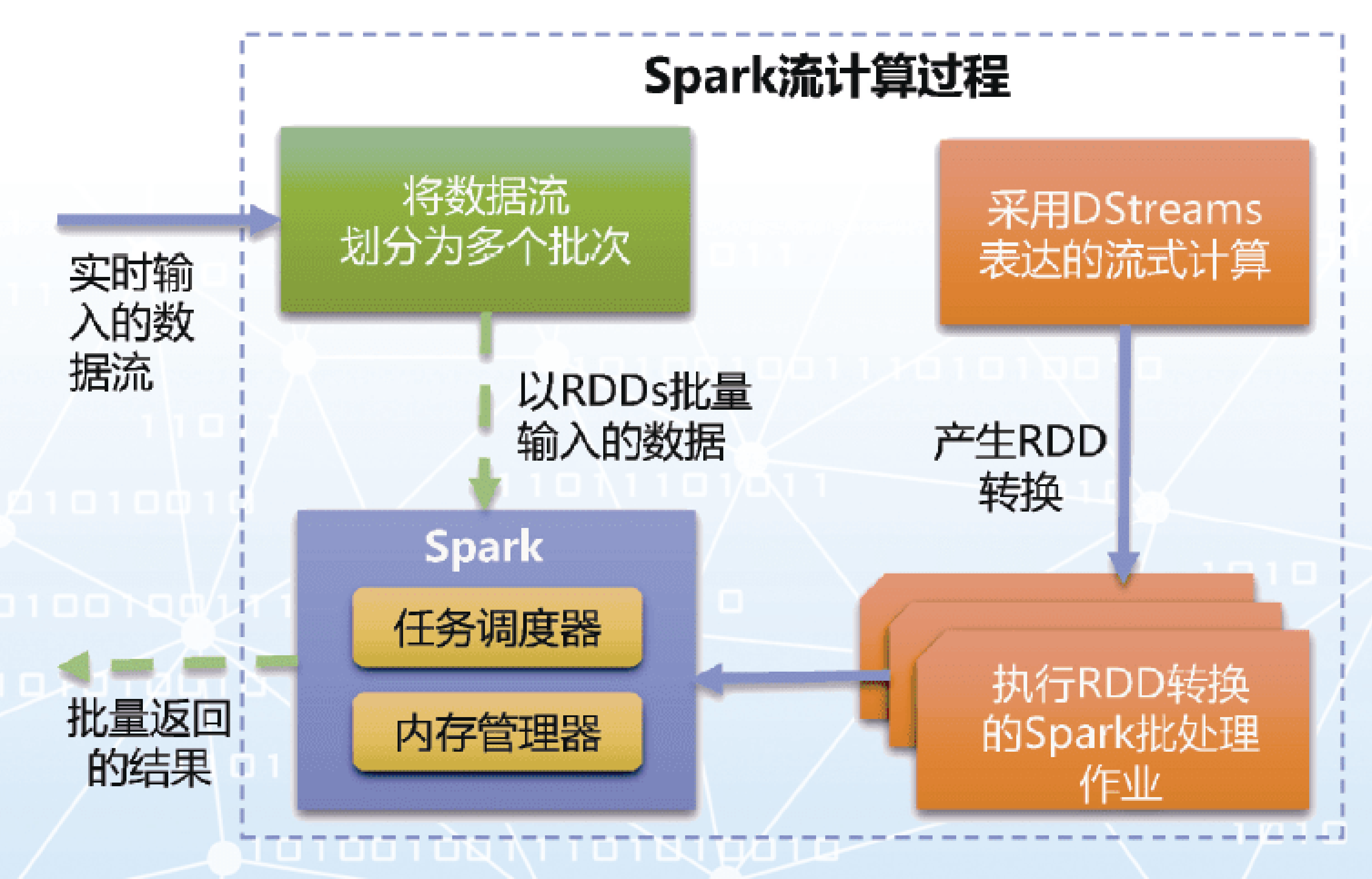
ITIL(v3)总体架构

服务设计
服务转化
服务战略
服务运营
服务持续优化
服务持续优化





内存计算框架Spark



献给

亲爱的妻子杨阳，生命因你而更加完善

献给

调皮捣蛋的儿子哲哲，生活因你而充满期待

致谢

首先感谢我的妻子杨阳，为了让我能够全身心地投入到写书工作中，主动为我承担了照顾家庭和孩子的事情；感谢我的父亲、姐姐和姐夫，帮助我照顾母亲，希望母亲能够早日康复；感谢我的哥哥和嫂子对本书写作一直的关心与支持；感谢伟大的游泳运动，让我在写作过程中能够一直保持健康的身体。

感谢邮电设计技术杂志社的曹军苗女士，曹老师在本书写作之初给予的热情帮助和鼓励令我终生难忘；感谢天文学家李宗伟教授对于图书写作一直的鼓励与支持；感谢中国联通研究院软件与系统实验室的罗云彬主任对本书写作的理解与支持；感谢 10 位行业专家在百忙之中审阅本书，反馈宝贵意见并提供推荐评语；感谢李景春先生、孙东光先生、张融女士以及老同学韩炳光为本书给予的热情支持与帮助。

感谢多年来一起工作过的领导和同事，虽然你们没有直接参与本书的编写工作，但是本书也蕴藏着与你们交流过程中产生的火花。最后，感谢那些辛勤分享知识的网友、图书著作者们，是因为站在你们这些巨人的肩膀上，才会有这本书的思考和新生。

专家推荐

1．数据的运营和管理将成为每一个企业都要面对的问题。这本书以企业架构为中心，阐述了如何把基于大数据的思维、方法和技术融合在企业的管理之中，服务企业的业务目标和活动。此外，作者还系统地梳理了当下信息技术发展中的各个热点，这也是本书的一大亮点。

阿里巴巴集团副总裁，著名信息管理专家，《数据之巅》作者 涂子沛

2．当前中国经济正处于从“要素驱动”向“效率驱动”的大转型时代，效率驱动的主体就是广大企业要从粗放增长向集约增长转型，通过技术创新、管理创新和制度创新不断提高生产力和竞争力水平，实现效率驱动经济增长的良性循环。看完本书，让我更加坚信大数据将是驱动这个时代转型的“新能源”、“新引擎”。作者利用完整人生的拟人手法系统、科学地提出的“全方位架构企业，赢在大数据运营”的全新观点，将使更多企业有所感悟、有所作为。

国家无线电监测中心、国家无线电频谱管理中心副主任，宋起柱 博士

3．本书作者将企业架构和大数据两个视角相互融合，提出构建面向大数据时代的服务型企业架构新思维，视角独特，理论新颖，对于企业形成大数据思维，设计大数据商业模式，构建大数据应用架构具有很好的参考价值。

北京赛智时代信息技术咨询有限公司总经理 赵刚 博士

4．大数据时代已经到来，但是许多企业对于大数据服务运营仍然比较陌生。对于许多企业而言，其掌握的数据实属金矿；然而，如何开采、管理、运营、变现这座大数据金矿是一个具有挑战性的课题。本书对大数据服务和运营做了系统、深入、翔实的讨论，其中不乏许多从实际应用中提炼出来的深刻见解。作者在技术管理、工程规划、软件开发等领域有着丰富的经验，把自己积累十几年的一线经验知识梳理出来，撰写出这本关于大数据运营的专著，非常及时，也非常值得一读。

华裔计算机科学家，美国华盛顿大学副教授、终身教授 陈一昕 博士

5．从企业理财的角度看，大数据运营将导致企业融资决策、投资决策、财务优化、税收筹划和风险治理的革命性变化，本书的优异之处在于为企业适应这种变化准备了技术、方法和实践。

财富管理专家 宋晓恒 博士

6．在信息膨胀的社会，无系统化的思维便会落后挨打。本书不絮叨大数据的好处，又不晦涩地仅谈技术，而是自业务说起，强调如何运用大数据搭建企业框架，这对于企业管理人员开拓思路，搭建工作平台均会提供一臂之力。

东方基金管理有限责任公司总经理 孙晔伟 博士

7．通过作者独具匠心的构思和设计，本书从企业架构和战略平滑地讨论到大数据运营，以及相关技术体系的建设和完善，既适用于企业管理人士，更适用于 IT 主管和 IT 设计及研发工程师，还可以帮助企业信息化部门的管理和技术人士更好地厘清业务与 IT 之间的关系，因为他们是大数据运营的主要承担者。

中讯邮电咨询设计院有限公司信息技术部总工程师 梅斌

8．本书从电信行业出发，全面系统地阐述了作者对企业架构、大数据运营的深入理解，既着眼于战略体系高度，又注重具体案例分析，是集作者多年规划、设计、管理经验的呕心沥血之作。

京东商城大数据专家 李净 博士

9．无论是企业大数据还是科学大数据，无论卖不卖钱，都需要“运营”。凡事预则立，不预则废。这是一本企业管理者和数据科学家可以一起分享的书。

国家天文台研究员，中国虚拟天文台和中国天文数据中心负责人 崔辰州 博士

10．大数据是一个系统工程，作者紧紧抓住“架构”、“过程”、“服务”三个要点，对大数据运营的方方面面娓娓道来，是目前为数不多的在行业大数据实操层面解析详尽，切实落地的专业著述。收到本书的时候我正在思考如何进行大数据公司的数据治理，感谢作者以其坚实的理论基础和丰富的实践经验，为大数据运营提供了一个周密的框架体系和可执行的实施路径。虽然本书主要围绕电信行业，但对于其他行业同样具有高度的参考价值。书中阐述的高屋建瓴的架构体系规划，业务驱动的大数据服务设计，组织严密的过程实施，相信无论是高层的战略规划者、中层的管理者，还是基层的执行者，都能从本书中获益匪浅。这是一本服务型企业在大数据时代进行价值创造的行动指南。

艾漫科技副总裁 郭锐

序：全方位架构企业，赢在大数据运营

科学技术的发展大大改变了人类生产和生活的方式，尤其是自从人类发明了计算机和互联网以来，信息的快速流动和共享让全球资源得以有效配置，有力地推动了世界经济的全球化和一体化。

信息通信技术的发展引发了多个社会热点，包括物联网、移动互联网、云计算、大数据等。物联网的目标是连接自然环境与物质世界，移动互联网的目标是连接人与人，云计算的目标是实现 IT 资源如同水电一样按需分配，大数据的目标则是为不同领域提供决策支持。

物联网能够实现物与物的连接，可以应用于工业、环保、医疗、交通、安防、水利、物流、仓储等领域。据 IT 研究与咨询公司高德纳（Gartner）预测，到 2020 年，可穿戴设备出货量将达到 5 亿，世界将安装 300 亿个无线传感器节点，未来 10 年传感器数量将会达到万亿级。预计 2020 年，亚太地区的物联网设备总量将从目前的 31 亿台增至 86 亿台，除日本外的市场规模将从目前的 2500 亿美元增至 5830 亿美元。中国将成为全球物联网市场的领跑者，预计 2020 年，中国物联网市场规模将占整个亚太地区市场规模的 59%。可见，物联网产业具有非常巨大的发展潜力。

物联网连接的是“物”，而互联网连接的则是“信息”。随着移动通信技术的发展，移动互联网已经成为社会发展的热点之一。根据中国互联网信息中心统计，截至 2014 年 6 月底，我国网民数量达到 6.32 亿，手机网民数量达到 5.27 亿，移动互联网用户大约占互联网用户总数的 80%。

目前，全球大概有 52 亿移动用户，其中仅有大约 30%的智能手机使用率，具有很大的市场发展空间。据美国权威市场研究公司 IDC 预测，2015 年全球智能手机出货量将达到 14 亿，市场规模将达到 4840 亿美元，中国智能手机的出货量将达到 5 亿部，超过全球出货量的 1/3。2014 年，中国移动市场规模接近 1900 亿元，预计 2015 年，中国移动互联网市场规模会超过 4000 亿元，预计 2017 年，移动互联网市场将继续保持强劲的增长势头，有望超过 6000 亿元。

从产品形态看，传统的智能手机、平板电脑将会逐渐向大屏、高清显示、多核处理器、多模多频的方向演进，而可穿戴设备、跨界智能终端、智能电视、智能汽车等将成为新兴的智能终端产品，具有非常强劲的市场潜力。

移动互联网市场在移动搜索、移动在线教育、移动电子商务、移动支付、移动在线游戏等方面将会保持强劲的发展势头。预计 2015 年，移动搜索的市场规模将超过 60 亿元，移动电子商务市场将突破千亿元，移动支付市场规模将超过 7000 亿元。

移动互联网的发展离不开移动智能终端和移动通信网络的发展，而移动通信技术是推动移动互联网飞速发展的前提和基础。作为移动互联网的重要载体，智能手机、平板电脑等移动设备的销量将继续扩大。除了智能手机和平板电脑，手表、手环、项链、滑板、智能眼镜、环境监测设备、医疗设备等将会有很大的发展空间。

在移动数据业务的支持方面，移动通信网络的发展经历了第二代（2G）、第三代（3G）和第四代（4G），第五代（5G）正处于研究阶段。2G 的传输速率为 9.6kb/s，最高可达到 384kb/s；3G 在室内、室外和行车环境的传输速率分别为 2Mb/s、384kb/s 和 144kb/s，通过优化最高可达到上行 5.8Mb/s，下行 28Mb/s 的传输速率；4G 的传输速率可以达到上行 50Mb/s，下行 100Mb/s；5G 的传输速率预计最高可达到 10Gb/s，是 4G 传输速率的近 100 倍。移动通信网络传输速率的不断提升为移动互联网应用的发展创造了条件。

云计算能够实现 IT 资源的按需分配，推动更加专业化的社会分工，进一步激发全社会的创新能力。据美国权威市场研究公司 IDC 预测，2015 年云计算的市场规模将达到 1180 亿美元。根据计世资讯研究（CCW Research），2014 年我国云服务市场规模已经达到 1645.8 亿元，同比增长 28%，其中，IaaS 占比达 23.4%，SaaS 服务占比约为 70%，PaaS 的市场占有率较低。为了推动云计算在我国的快速发展，工业和信息化部在“十三五”纲要中，将云计算列为 2016—2020 年重点发展的战略性产业。

据美国权威市场研究公司 IDC 预测，2015 年大数据相关的软硬件及服务市场规模将达到 1250 亿美元，图像、音频、视频等多媒体成为大数据分析的重要驱动力，将会呈现至少 3 倍的增长。产业链中起主导作用的 IT 服务提供商将提供数据即服务（DaaS）平台，大数据分析公司会在此基础上提供增值服务，物联网领域将成为主要的分析对象，预计在未来的 5 年，大数据市场将会有 30%复合增长率的高速增长。

无论是物联网、移动互联网，还是云计算、大数据，都预示着未来巨大的市场发展空间。作为社会经济细胞的企业，在面对市场提供的各种发展机遇时，需要根据自身情况制定发展战略，在激烈的市场竞争中占得先机。

信息通信技术、交通技术促进了全球在投融资、设计、采购、生产加工、物流配送、渠道销售等环节更加专业化的社会分工，大大提高了社会整体效率，推动了人类社会的快速发展。

社会生产在全球范围内的分工，虽然促进了社会生产力的发展，但是也使得企业处于风险更大的环境之中，企业需要快速地响应外部变化，才能够在市场竞争中占据主动。因此，要求企业能够将发展战略有效地贯彻到建设和运营活动之中，提高执行力。如果企业发展战略与建设和运营脱节，企业将会偏离预先设定的目标和方向，在激烈的市场竞争中处于不利地位，甚至会破产。因此，企业应当从多维度、全方位地架构企业，确保企业发展战略能够真正落地实施。

企业架构是复杂的系统工程。企业通常需要定期进行外部环境分析和内部资源评估，制定中长期发展战略。因此，要求企业架构模型能够完成对目标和现状的分解要求，通过差距分析，为企业制订明确的行动计划。

企业的业务活动往往需要不同部门、不同角色、不同地域的人员共同参与，比如市场营销部门的营销人员、产品销售部门的销售人员、客户服务部门的客服人员、采购部门的采购人员、工程建设部门的建设人员、维护部门的维护人员等。此外，人力、财务、资产等职能部门也需要共同参与，如果没有良好的企业架构，则很难保证企业发展战略能够有效地贯彻到企业建设与运营活动之中。可见，企业架构在企业中的重要地位。

物联网、移动互联网、云计算等产业的发展，为全社会生产了越来越多的数据，为了体现这些数据的新特征，业界将其定义为“大数据”。如果说煤、石油、天然气等是自然界提供的能源，那么大数据则是信息社会提供的新型能源。对于企业而言，大数据成为企业认识市场、客户和自身的核心资产。

大数据虽然魅力无穷，但是如果企业不能正确认识和利用大数据，那么对于企业而言，数据仍然是一堆废铜烂铁。关于大数据，企业需要引发许多思考，包括大数据如何在业务活动中发挥作用？如何发现和定义大数据服务？如何设计大数据服务？如何部署大数据服务？如何持续地运营大数据服务？如何有效地管理大数据？大数据服务在企业架构中如何承载？等等。

可见，对于企业来讲，要想充分理解和运用大数据服务，并不是一件非常简单的事情。为了便于读者快速地掌握本书中关于大数据运营的思路和方法，笔者在此概括性地说明一下。

大数据来源于自然环境和人类社会，是对自然环境、人类特征和行为的记录，其原理

是借助数据来把握规律，进而实现预测未来和支持决策的目的。因此，可以将业务活动分为两类：负责执行的业务活动和支持决策的业务活动，大数据服务属于支持决策的业务活动。

企业架构框架从空间角度来架构设计企业，由于决策和执行是业务活动，是一体两面、不可分割的，因此大数据服务与面向操作的事务型活动一样，在企业架构的10个维度（业务过程、信息、应用、功能、数据、集成、技术、部署、安全、治理）需要相互配合，共同支撑完成从企业发展战略到运营的转换。

大数据服务在时间维度上体现为从需求分析、架构设计、开发测试、部署上线到持续优化的过程。不同于面向操作的事务型应用，大数据服务需求来源于业务需求和大数据两者的结合，业务需求是待决的决策问题，而大数据则是解决问题的数据基础。在架构设计方面，大数据服务重点关注数据的全生命周期管理和元数据管理，大数据是长期历史数据的积累，应当根据应用需求和管理要求制定数据迁移策略，元数据相当于分析人员数据字典。在持续优化阶段，企业需要借助监控手段，实时监控大数据的活性和运行状况，不断丰富和完善数据源，提升数据质量，不断提升决策支持的及时性和正确性。

本书是笔者的处女作，由于个人认识水平和时间的限制，不足之处恳请批评指正。希望本书能够让读者更加全面、系统地掌握基于大数据架构企业的思路与方法，充分挖掘大数据资源的潜力，赢在充满希望的大数据时代！

个人联系方式：

邮箱：lifudong00@tsinghua.org.cn

微信：dsjlfd，QQ：371574651，QQ 群：156966796

博客：lifudong.blog.51cto.com

个人网站：www.easyarch.cn

李福东，2015 年 4 月于北京

目录

导读

从企业战略制定到战略实施是一个复杂的过程，那些高大上的咨询成果难以落地成为困扰众多企业管理者的难题。因此，从企业战略出发，多层次、多维度、体系化地设计企业架构，成为保障企业战略落地的有效手段。企业架构有效衔接了企业发展战略、基础设施建设、生产运营、企业管理等多个环节，并以服务能力为中介，有效地衔接了业务与技术，成为企业战略落地实施的指南针。

大数据时代的到来，为企业提供了另外一种提升核心竞争力的方式和手段。企业可以通过吸纳、整合和挖掘隐藏在大数据背后的规律，理解和把握客户需求，按需生产，同时运用互联网思维，有效匹配能力需求和资源供给，以最具有成本效益的方式提供产品和服务。

大数据不能脱离企业业务活动而单独发挥作用，企业应当基于企业架构完成大数据服务的设计和运营，这样才能让大数据找到立足点和归宿。以电信运营商为例，每月都会产生 PB 数量级的通话行为和上网行为记录，如此大量的通信业务使用记录中蕴藏着巨大的能量和价值，人们既可以基于通信行为分析用户偏好，实现精准化营销，也可以结合电信运营商的基站使用情况、用户价值、应用价值等辅助完成无线网络的规划设计。当然，通信大数据也可以作为资产对外销售，更大程度地发挥通信大数据的价值。从某种程度上说，大数据的潜力取决于企业的想象力。

随着企业对大数据理解的日益深入，大数据服务的数量势必会不断增多，有效管理和发现新的大数据服务逐渐成为挑战性的难题。一方面，人们可以从企业业务活动出发，分析企业业务活动中的决策环节需要什么决策输入，另一方面，也可以从大数据出发，分析大数据的决策支持能力，两种方法合在一起就能更加快速地发现新的大数据服务。这样不但解决了大数据服务难以管理的问题，而且可以做到有的放矢，挖掘出更多的大数据服务。

企业业务活动可以分为两大类。一类属于操作型活动，例如信息维护、订单提交、工单流转等，这类活动的关键在于“行”，追求执行的“正确”；另一类属于分析型活动，例如客户偏好分析、资源消耗分析、财务分析等，这类活动的关键在于“思”，追求“正确”的执行。大数据用于支持分析型活动，保证操作型活动能够“正确”执行。

要使大数据能够更好地支持企业决策活动，需要企业具备良好的大数据运营能力，包括获取更多的数据源，保证数据的准确性、及时性，等等。可见，大数据服务从创意到形成并非一朝一夕的，是一个长期的、不断探索的过程，企业只有掌握运营大数据的方法，才能够发挥大数据的价值。

为了展现从企业战略到大数据运营的全过程，本书按照一个人从成家立业到奋斗腾飞的过程，将内容分为筑巢、联姻、孕育、分娩、培育、腾飞 6 个阶段，像经营家庭一样经营大数据。

第一阶段为“筑巢”阶段，目标是完成企业的整体架构设计。为了全面、系统化地展现企业架构模型，本书从企业战略角度出发，从 10 个视角对企业进行了架构设计，这 10 个视角既相互区别又相互联系，宛似一座小房子，因此取名为“筑巢”，意味着企业要像构筑一座房子那样严谨，以便让企业的各个部分能够协同配合，发挥合力。

第二阶段为“联姻”阶段，目的是解决大数据与企业架构结合问题。大数据只有与企业业务活动有机结合才能发挥作用，业务活动是企业架构的业务输入，业务活动中既有“执行”又有“决策”，大数据的作用就是帮助企业快速、准确地完成“决策”。大数据与企业架构的结合是一体化、不可分割的过程，就好比人们现实生活中的婚姻，男女双方在构筑更加美好的生活过程中都有贡献，缺一不可，因此第二阶段取名为“联姻”。

第三阶段为“孕育”阶段，目的是解决大数据服务孵化问题。操作型应用通常是先有需求后有数据，而大数据应用则是先有数据后有需求，两者正好相反。因此，对大数据服务的需求分析和设计与操作型应用采用不同的思路和方法。大数据服务设计是一个从大数据能力朝着待解决问题不断靠近、反复迭代的过程，整个过程漫长而充满期待，好比精子寻找卵子一样充满风险和挑战，因此将这一阶段称为“孕育”。

第四阶段为“分娩”阶段，是大数据服务设计和开发成果向大数据运营转换的阶段。不同于操作型应用的转换过程，这个阶段会经过多次调整和完善，直至达到最满意的答案，例如数据清洗、数据转换、数据装载、数据稽核、模型调整、算法优化等。此阶段类似于十月怀胎后的分娩过程，也许十分顺利，也许要经历多次镇痛，但是只要保持耐心，随机应变，终究会得到满意的结果。

第五阶段为“培育”阶段。大数据服务并不是一劳永逸的，企业外部竞争环境总是不断变化的，大数据的数据源、数据时效性等也在不断发生变化，这些都需要企业重新审视大数据服务，重新调整分析模型和算法，这好比家庭培育孩子，总是需要根据社会要求来调整培育方向，使得孩子更能够适应社会发展的需要。这一阶段取名为“培育”，意思是对大数据服务的“培养和教育”。

第六阶段为“腾飞”阶段。在这一阶段，大数据的价值完全得以体现，可以说，经过前面几个艰难困苦的阶段，大数据服务终于可以扬眉吐气了。大数据魅力无穷，取得了一个又一个辉煌的成果，因此本阶段取名为“腾飞”。

此外，本书内容的体系结构和方法论主要参考了 Frameworx/NGOSS、ITIL/ITSM 两个国际标准规范。Frameworx/NGOSS 方法论解决业务活动与大数据的结合问题，ITIL/ITSM 方法论则侧重解决大数据服务从设计、转换到运营的衔接问题。

大数据技术是保障大数据应用落地的重要手段，在本书最后一部分专门分析大数据相关技术，如 Hadoop、Oracle、R、GIS、Android 等的原理和方法，使得大数据运营体系更加完整。

本书中提出的大数据运营方法论以支持服务型企业为主，同样也可以作为政府机关、事业单位、科研院所等构建和运营大数据服务的参考资料，具有良好的行业通用性。本书主要以电信、金融和互联网 3 个行业为主线进行分析。

本书面向的读者对象不但包括企业战略层面的管理人员，如企业的总经理（CEO）、信息总经理（CIO）、技术总监（CTO）等，同样适用于项目经理、系统架构师、数据分析师、开发工程师、测试工程师等掌握大数据运营过程的人员。

引言

信息通信技术、交通技术的发展促进了经济的全球化和一体化，信息的自由流动实现了各种资源在全社会范围内的配置，社会专业化分工更加细致，社会经济更加具有效率和活力。

科学技术在推动社会发展的同时，也让企业处于更加不确定的经营环境之中。企业需要具备敏捷地响应变化的能力，需要解决好发展战略到日常运营的过渡问题，需要解决好业务与技术的衔接问题，最终形成一个环境自适应的、能力不断优化完善的管理体系。

企业要解决好以上问题,首先需要具备以下思维方式，实施系统化的架构设计，主要包括 5 个方面：

第一，要认识到商业模式已经从价值链条转变为价值网络模式。价值网络模式要求企业能够在社会分工中把握好适合自身发展的关键环节，具备良好的集成能力，实现业务能力的组件化和服务化。

第二，要认识到 IT 架构模式已经从面向单一系统转变为面向服务的模式。竖井式的系统设计使得组织业务流程流转不畅、信息难以充分共享。面向服务的架构模式将业务能力和 IT 能力视为一种服务，使得企业内部和外部均可以通过服务的方式进行交互。

第三，要认识到数据是推动企业发展的核心资产。与传统的资产不同，数据可以帮助企业及时、准确地认识市场、客户、供应商、合作伙伴、员工等的需求并采取适当的行动，可以说，大数据是企业未来发展的生命线。

第四，要正确认识操作活动和分析活动之间的密切关系。操作活动好比人的四肢，主要负责执行，而分析活动则好比人的大脑，负责思考和决策。正确认识两者之间的关系，可以使企业从业务活动的角度出发，将两类活动连接起来。

第五，要正确认识职能、过程以及全生命周期管理之间的关系。职能管理面向企业某一特定功能，过程管理采用业务活动分类方法，将企业业务活动分解为多个相互配合的过程块，过程块之间相互配合实现不同的职能。全生命周期管理要求从事物产生、发展、消退、消亡的全过程思考问题，让认识更加全面。

有了设计良好的架构，企业就具备了连接战略与运营、业务与技术的桥梁和纽带，才能够将发展战略有效贯通到企业的日常运营活动之中，同时也能够实现业务需求与技术支撑的无缝对接。

大数据服务与操作型服务相比，既有自身的独特之处，又有着密切的联系。

第一，大数据服务的目标是支持决策的制定，而操作型服务用于支持业务操作的完成。

第二，与操作型服务相比，大数据服务对于系统的响应性要求较低，操作型服务对响应时间通常要求在秒级。

第三，大数据服务的数据操作主要是读操作，而操作型服务主要为写操作，要求事务必须是完整的。数据存取特点不同，数据架构方案也不同。

第四，大数据服务依赖的数据规模大而且数据量会不断增加，要求存储架构具有良好的线性扩展能力，通过横向基础设施的扩展，就可以实现数据存取能力的线性提升。

第五，大数据服务更像是一个探索发现的过程，大数据服务需要持续提升数据的完整性和准确性，而操作型服务则更关注于对业务需求的满足、易用性以及操作效率。

大数据服务与操作型服务也有着密切的联系。

第一，大数据服务的数据源头是业务操作和业务使用日志，无论这些数据是企业内部应用产生的还是其他组织产生的。

第二，大数据服务与操作型服务业务活动是一体的，不可分割。大数据服务负责分析判断，而操作型服务则负责执行，两者是“知”与“行”的关系。

第三，大数据服务与操作型服务都支持战略、战术、执行 3 个层次的业务活动。高层级业务活动重点在于确定方向和路线，要求大数据服务能够提供全面、准确的分析结果，而低层级业务活动重点则在于执行效率，要求大数据服务能够快速地反馈分析结果。

在企业的各种业务活动中，虽然大数据服务与操作型服务起的作用不同，但是两者的实现思路却是非常相似的，都需要经历需求分析、架构设计、功能开发、测试部署、运行维护、优化完善、管理治理的过程。

企业架构可以衔接发展战略和日常运营，从整个业务活动的角度，大数据服务与操作型服务是不可分割的，因此大数据服务同操作型服务一样，需要从企业架构的 10 个视角进行分析、设计、开发、测试以及管理。

操作型服务的需求分析与设计输入是企业提出的业务需求，而大数据服务在需求分析与设计方面的输入则是待解决的决策问题和大数据基础。

当大数据服务按照要求完成设计和开发工作后，同样需要从开发测试阶段转换到上线

运营阶段，正式支持企业的生产经营。

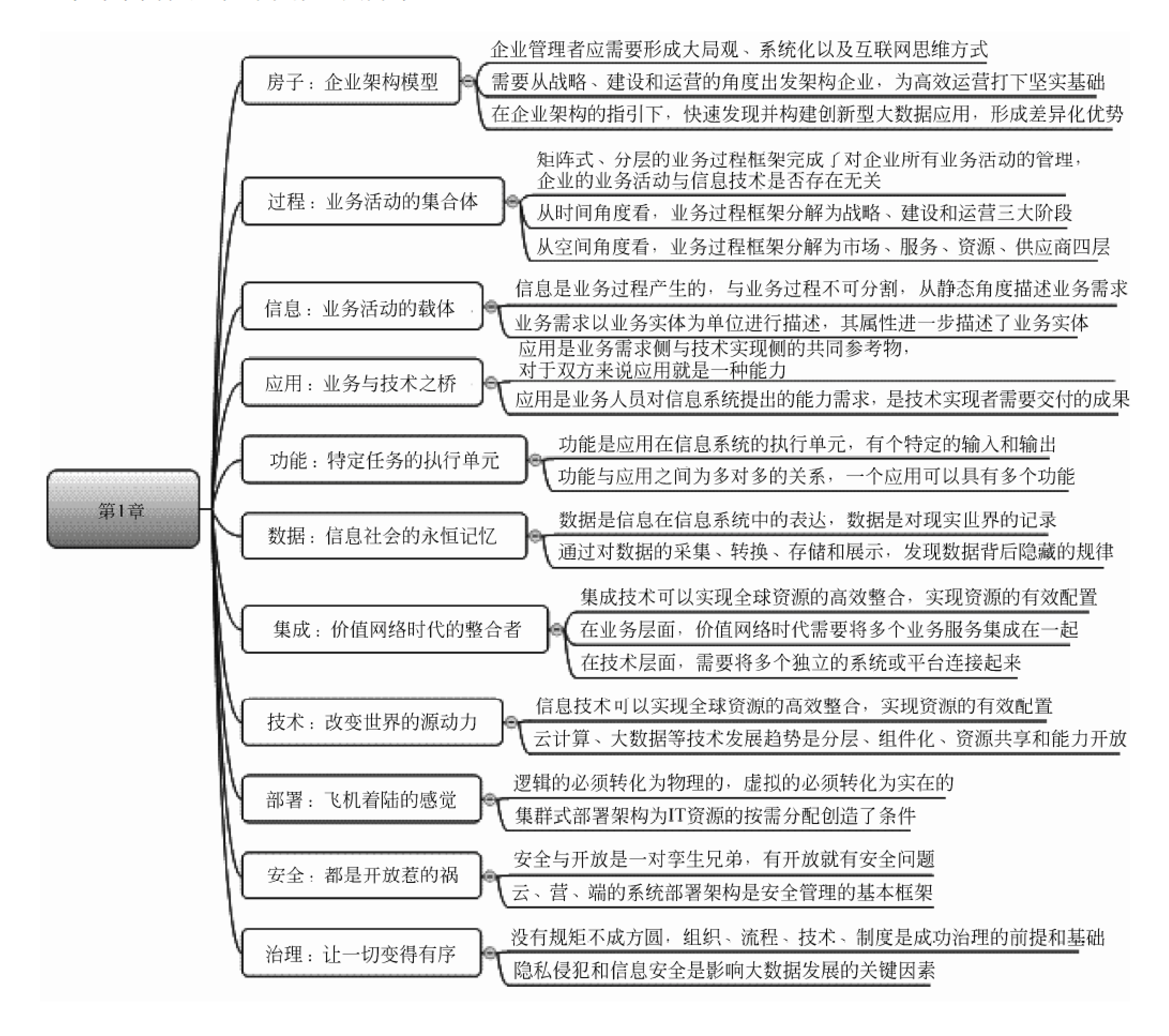
处在上线运营状态的大数据服务并不能一劳永逸，需要进行不断的优化完善。通过对大数据服务在数据采集、集成、清洗、转换、装载等不同阶段的观察，改善数据质量，提升系统的可靠性、可用性和性能，根据数据的活跃度和管理要求采取不同的数据迁移策略。

下面以筑巢、联姻、孕育、分娩、培育、腾飞 6 个阶段为主线，分别论述大数据服务在企业架构设计、与企业架构结合、需求分析与架构设计、转换、持续运营、行业实践的方法与思路。

第 1 章
筑巢：来自建筑行业的启示

大数据对于企业非常重要，但是如果没有设计良好的企业架构，很难看清楚它对于企业的价值和作用，因此本篇是在分析大数据运营之前的必要准备，通过企业架构的设计，可以清晰地看到大数据在企业中的发力点，进而形成满足大数据运营的企业架构新思维。

本章内容思维导图如下所示：



1.1 谋划：像盖房子一样架构企业

以企业发展战略为指导，结合业务架构与技术架构，按照系统的方法论，将企业架构绘制成一座 10 个维度的小房子。

由于企业自身的复杂性以及外部环境对企业响应要求的敏捷性，要求企业从战略、业务和技术方面统筹考虑，有效衔接。

企业架构需要从战略、业务、技术三个层次进行分析与设计。战略、业务、技术是三个相互联系又相互区别的部分，负责企业架构的人员应当从上到下、从前到后、系统化地对企业进行分析和设计，从而保证企业架构的整体性。

企业战略侧重关注企业发展的长远和全局，通过分析自身特点和外部环境，找出自身的优势与不足，同时确定机会和威胁，在知己知彼后制定符合企业自身的发展战略。例如，某电信运营商通过分析认为，3G 市场竞争中自身拥有的 WCDMA 技术相比其他竞争对手具有技术先进、产业链条更加成熟等优势，但在 2G 网络（GSM）的覆盖规模、网络质量等方面均与竞争对手相比存在较大差距，因此制定了 3G 领先战略，借助 3G 优势取得领先优势。

业务是在企业战略的指导下完成的，同时业务也需要技术的支持来实现。从企业内部看，需要整合内部和外部资源为客户提供服务。业务除了包括面向客户的市场营销、销售、服务以及支撑业务运营的客户、产品、渠道、合作伙伴等元素外，还包括面向企业内部管理的人力资源、财务、资产、工程、知识、风险等方面。此外，支撑企业提供服务的资源则是企业价值提供的基础，比如电信运营商的通信网络资源，银行的货币资源、电力公司的电网资源等。

技术是一种手段，用于支撑业务需求的实现。在信息、通信、物联网等技术飞速发展的时代，技术在提升运营效率、管理水平、客户感知等方面都发挥了越来越重要的作用，同样，技术应当与业务紧密关联，应当能够迅速响应业务需求的变化。

企业架构应当能够紧密连接战略、业务与技术，作为指导企业发展和响应外部变化的蓝图。企业应该多个维度定义相互联系、相互制约的架构蓝图，用于评估企业发展现状和目标，通过目标与现状的对比，企业从多个维度定位问题和差距所在，并根据改进的原则、

方法和工具进行不断调整，以实现企业整体目标。

本章按照从战略到业务再到技术的思路，以企业发展战略为指导，结合业务架构与技术架构，按照系统的方法论，将企业架构绘制成一座小房子，如图 1-1-1 所示。

从这座房子可以看出，企业架构共从 10 个维度进行管理，分别为：业务过程架构、信息架构、应用架构、集成架构、功能架构、数据架构、技术架构、部署架构、安全架构和治理架构。其中业务过程架构和信息架构属于业务层面，功能架构、数据架构、技术架构、部署架构属于技术层面，应用架构和集成架构处于业务和技术的衔接点，起到桥梁和纽带的作用，而安全架构和治理架构则属于管理层面。

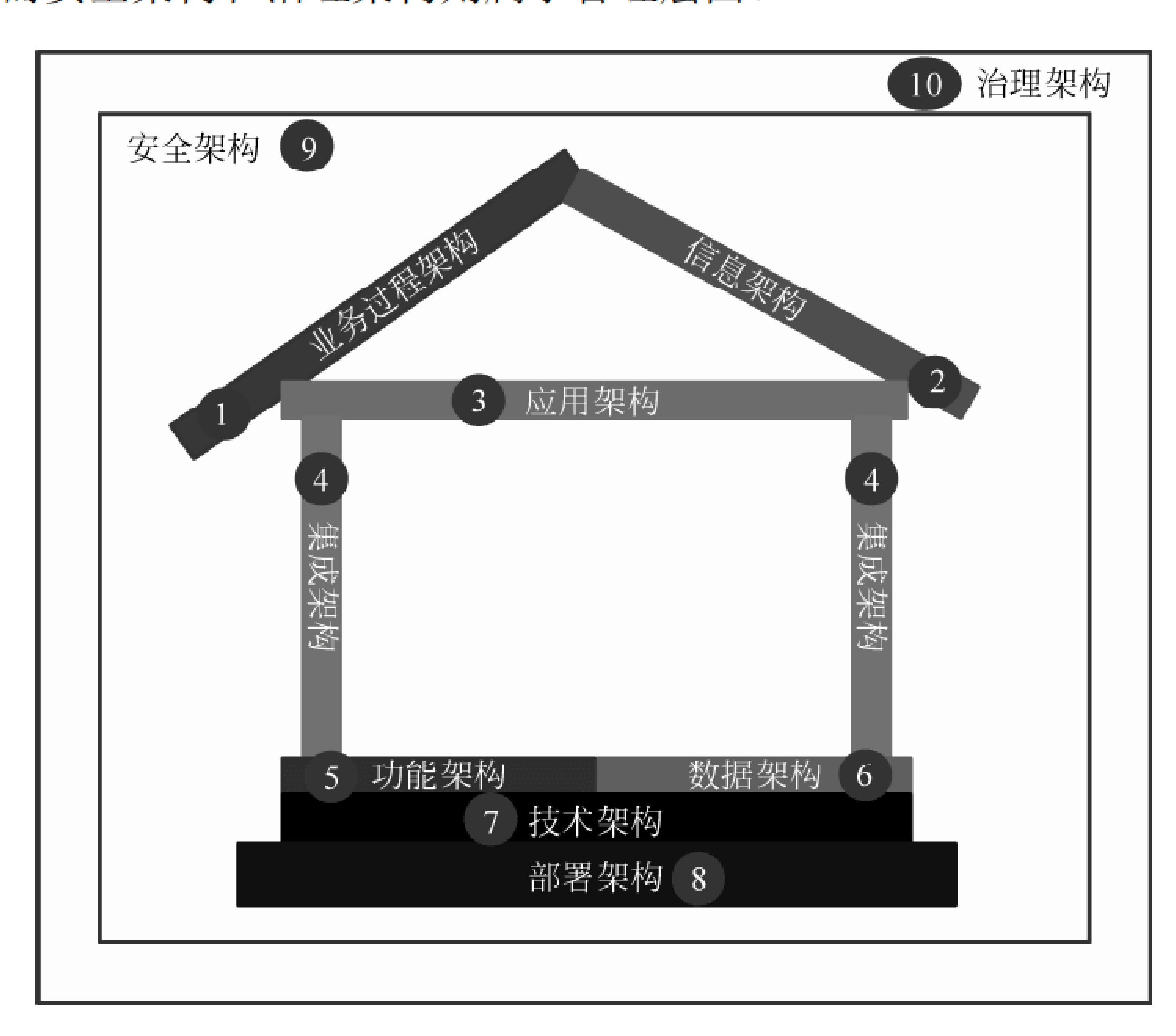


图 1-1-1　企业总体架构模型

下面就对这 10 个架构进行简要说明。

（1）业务过程架构和信息架构属于业务层面，用于描述业务流程、业务规则、信息规则等。

（2）信息架构：信息架构和业务过程架构仿佛夫妻关系，如果说业务过程架构从动态角度管理企业的业务活动，那么信息架构则是从静态角度管理企业业务过程中产生的信息的。

（3）应用架构属于业务与技术的过渡层，通过能力的形式体现业务需求，同样也是对

系统实现提出的要求，是业务人员和技术人员之间的共同约定。

（4）集成架构的目标是实现对业务、应用、功能、数据、技术的黏合。从业务层面看，它集成了业务过程与信息；从应用层面看，它通过组合业务服务形成各种能力；从功能、数据和技术的层面看，它是系统功能、数据模型以及技术实现的黏合体。

（5）功能架构是从系统用户角度看，系统用户功能点的集合。功能形成的基础是应用需求或者能力要求，能力与功能是多对多的关系，一个能力可能需要功能点来实现，一个功能点也可能对应多个能力要求。

（6）数据架构为功能架构的基石。按照数据支持目标，分为面向应用的交易型数据和面向决策的分析型数据；按照从业务到技术逐步落地的思路，分为概念模型、逻辑模型和物理模型。

（7）技术架构是根据应用特点和功能要求，采用先进适用技术而设计的，需要考虑性能满足能力、技术成熟度、可移植性、开发者群体规模、实施难度等因素，遵循分层、模块化、组件化、开放性的设计原则。

（8）部署架构是应用软件、系统软件、系统硬件（主机、存储、网络等）的连接方式，部署架构的目标是保证系统可靠性、可用性、可伸缩性、高性能以及安全性，因此一般采用集群方式来实现应用部署。

（9）安全架构是为了保证系统在信息交换过程中的安全性，包括网络安全、信息安全、访问安全、安全管控等。需要构建一个从技术到管理的安全管控体系，实现对安全威胁的预防、发现、处理、分析等全过程、全方位的管理。

（10）治理架构是为了保证从业务到技术的全过程的规范性、有效性、严谨性，通过建立从业务到技术的服务支撑体系，并借助规范化的管理流程和保障制度，保障业务运营的连续性和系统运行的稳定性。

1.2 过程：企业是业务活动的集合体

按照分层分类的方法，从战略、建设、产品到运营的时间维和从市场需求到资源供给的空间维进行设计，业务过程框架表现为时空交叉的矩阵形式。

正如恩格斯所说“世界不是既成事物的集合体，而是过程的集合体”，企业的生产经

营活动也同样由大大小小的过程组成。本节参考国际先进成熟的框架体系，按照分层分类的方法，分别从时间角度和空间角度对企业业务过程进行架构设计。

1.2.1 分层分级——最原始的方法论

将复杂问题简单化的有效方法是分而治之（divide and conquer），对于企业来讲，通常是从供应商/合作伙伴那里获取原材料，通过企业自身的生产与运营，将产品和服务交付给客户，此外，企业的活动要受到政府等监管部门的约束，符合社会利益。虽然企业采购与生产经营活动看似简单，但是对于专业化分工越来越细、外部环境变化越来越快的今天来说，如果不借助科学的架构对企业进行管理，那么当企业面临问题和外部挑战时，难以定位问题所在。

那么如何对企业进行架构设计呢？首先就要分析一下企业的业务活动。从时间轴看，一个企业一般会经过战略制定、基础设施建设、产品管理、市场营销、销售以及客户服务几个阶段。从空间轴看，企业首先从供应商/合作伙伴处取得产品和服务，作为生产的原材料，需要对获得的资源进行配置，这些资源是承载客户产品和服务的基础，可以采用面向服务架构（SOA）的方式，将其进行封装，成为一个个前台可以使用的服务，通过服务的组合满足业务需要。但是，到现在为止，还不能直接将服务交付给客户，因为还缺少对于市场的支持，比如产品的结构怎样，如何定价，产品面向的客户群是谁，产品销售的渠道是哪种类型，是实体营业厅还是网上电子渠道等，因此需要在服务层之上再增加一层，这一层包括市场营销、产品、客户几个方面。最后，企业要完成以上业务活动，还需要企业人力资源、财务、采购、资产、研发等过程的支持。

以上是对于企业活动的一个简单分析，不难看出一个企业尤其是大中型企业并不是想象的那么简单，它是由许许多多过程共同来完成的。为了清晰地看到企业的过程，将企业从零级开始，按照分层的方式，逐步分类剥皮，最后达到最底层的执行节点。企业架构模型的零级视图如图 1-2-1 所示。

从图 1-2-1 可以看出，企业主要有三类利益相关者。第一类是客户，这是企业产品和服务的输出对象，客户可以有多种分类方法，比如可以分为政企客户和公众客户。第二类是供应商和合作伙伴，这是企业生产和运营的输入对象，包括设备供应商、软件提供商、系统集成商、内容提供商、服务提供商等角色。第三类是企业内部服务对象，包括股东、雇员、政府监管机构等，股东是企业的投资者和受益者，雇员是企业的经营者，政府监管

机构是保障企业符合市场要求和合法经营的管理者。

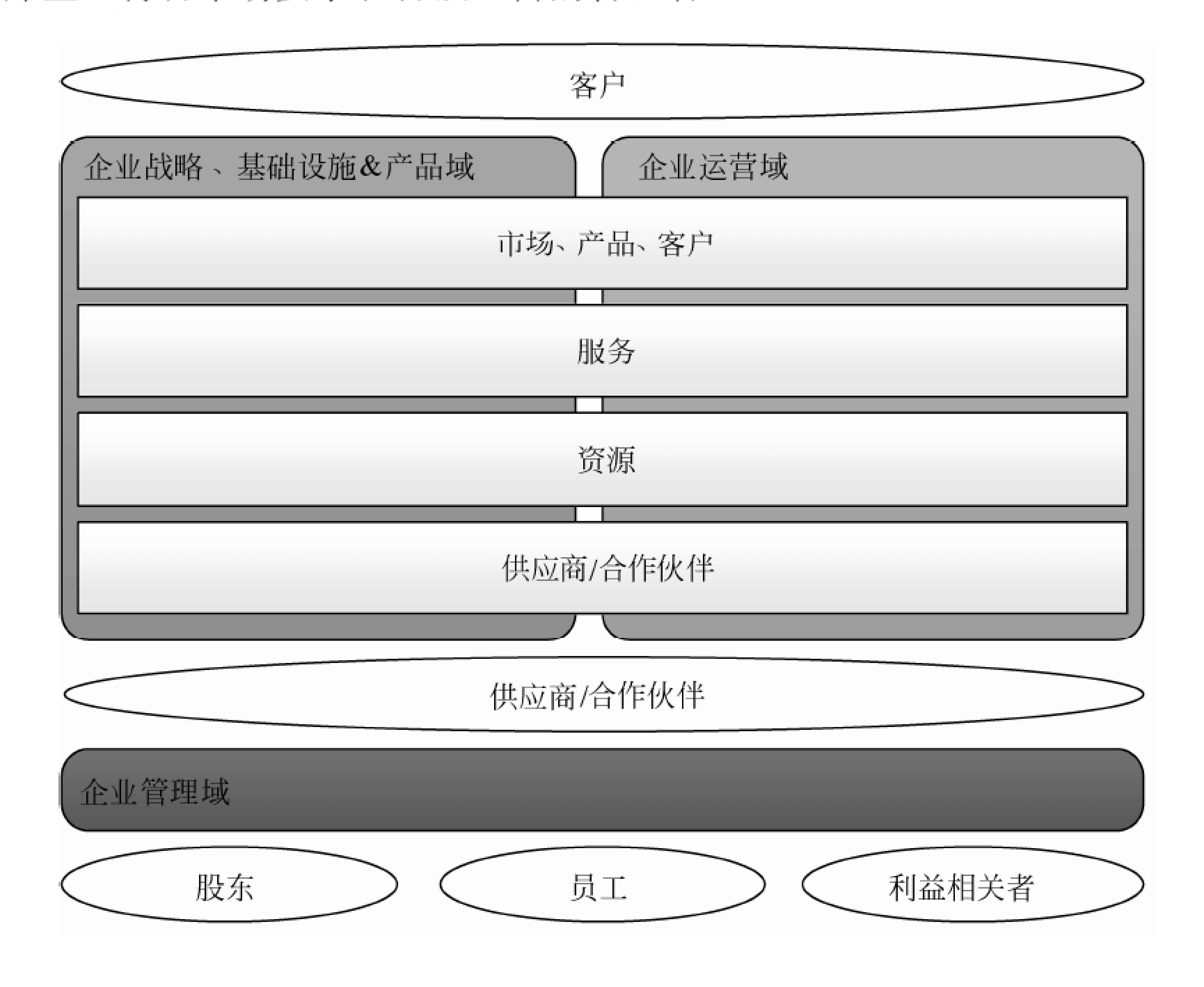


图 1-2-1　企业零级概念模型

企业的零级概念模型只是企业架构最高层次的抽象和分类，如果要管理好企业的过程还需要在零级概念模型上进一步剥皮。对于分层深度没有特别要求，原则上是能够将企业业务过程分为一个有特定功能的独立单元，使得企业架构中的各个元素之间保持松耦合关系。为了清晰地掌握企业架构分层的方法，再看一下企业的一级架构模型，如图 1-2-2 所示。

从图 1-2-2 可以看出，企业业务过程框架以客户为中心设计，解决了从市场需求到资源供给的承接问题，其原理为：首先，企业需要确定市场的 4P（产品、价格、促销、渠道）要素，进行客户关系管理，这些属于市场（客户）层面。其次，企业需要将市场需求转换为服务能力，比如某客户订购了一个电信融合产品，这款产品包含固话、宽带、移动三种通信能力，这些在服务层实现。然后，这些服务能力是虚拟的、逻辑的，它们需要企业真实的、物理的资源提供支持，比如，宽带服务能力需要交接箱、分线盒、光缆等线路资源的支持，这些在资源层实现。最后，企业的各种资源不一定是自身提供的，还可能由外部供应商提供，在某些情况下也可能需要租赁合作伙伴的资源和服务，以满足建设工期或者

成本效益等要求，这些在供应商/合作伙伴层实现。至此，按照市场、服务、资源、供应商/合作伙伴的分层方法，就完成了从市场需求到资源供给的映射，这种分层方法对于服务型企业是通用的。



图 1-2-2　企业业务过程框架（一级）

当然，以上是从空间角度实现了从市场需求到资源供给的映射，毕竟企业不是静止不动的，还需要从动态角度定义企业业务过程。过程描述事物如何变化，结构则描述了事物如何相互联系，为了全面地认识事物发生发展的全过程，需要采用全生命周期管理的思维方式。

笔者将企业运营前的过程分为企业战略管理、基础设施生命周期管理和产品生命周期管理三大阶段。企业战略作为指导作用，决定了企业建设和运营的方向和重点，比如企业开发面向新的客户群的新产品、在新的地域开辟新的市场或者建设网上渠道销售产品等，这些战略对于企业建设和运营都有影响，企业战略分为战略制定、战略实施和战略评估三个阶段；基础设施生命周期以企业发展战略为指导，定义市场营销和产品能力需求并按照

这些能力需求进行建设实施，对于不符合企业发展的基础设施，经过评估后下线；产品生命周期过程主要包括产品的开发与退出、产品营销传播及促销以及销售开发，此过程是在基础设施能力具备的前提下实现的。这个道理很简单，企业在运营之前怎么会没有产品呢？不但企业要具备产品，而且还需要对产品进行营销推广和销售开发，以便潜在客户和现有客户能够知道企业的新产品，企业还需要定义产品的补贴机会、开展销售相关的培训、制定潜在客户识别方法、制作产品销售过程和步骤等。

此外，企业对人力资源、财务、资产、知识等的管理也非常重要，笔者将其定义为企业管理域。企业管理域是企业战略管理、基础设施生命周期管理、产品生命周期管理和企业运营活动的大后方，主要包括人力资源管理、财务与资产管理、企业效益管理、风险管理等。

从上面的分析可以看出，业务过程框架是时间维（从战略、建设、产品到运营）与空间维（从市场需求到资源供给）的结合体，表现为时空交叉的矩阵形式。

采用这种分层方法，企业业务过程继续细分，直到过程元素（叶子节点）。业务过程细分的目的是更清晰地展现业务过程交互的细节。业务过程作为企业活动的起始，作为其他维度企业架构设计的输入。

1.2.2 CXO 的那些事儿：企业发展战略

企业所有的业务活动中，企业战略是第一个业务过程，它决定了企业发展的方向和道路，对其他过程起着指导作用。

许多人认为企业发展战略是企业高层管理人员的事情，对此，国际著名战略学家戴维（Fred R . David）提出了系统化的看法，他认为企业战略的主要任务是沟通，如果没有企业内部的充分沟通，进而达到对企业发展战略的理解，那么企业战略的实施将是一件很困难的事情。

企业发展战略立足于全局和长远，主要目的是确定公司中长期发展方向、策略等。制定企业发展战略的方法主要是 SWOT 分析法，其本质是“知己、知彼”，企业要掌握企业自身的优势和不足，也要分析企业外部的机会和威胁，根据分析结果确定企业发展战略。比如，企业在渠道、技术、产品方面和竞争对手相比具有渠道覆盖广、技术更加先进，同时产品也具有价格优势，但是不足的是企业在资金支持方面不及竞争对手，经常因为客户

回款慢而导致现金流中断，财务风险高。

在外边环境分析方面，通常采用 PEST 分析法，即政治与法律、经济、社会与文化、技术四个方面。比如新能源汽车，虽然在技术上比较先进，但如果在国家政策和法律支持方面还处于空白阶段，那么政策风险高。在经济方面，需要结合目标区域的居民收入来确定合适的销售策略，与当地居民的收入匹配起来。在社会与文化方面，要结合当地的风俗习惯来制定营销策略，比如我国的北方，人们通常长得高大，生活习惯较为粗放，可以考虑提供耐用、宽敞的汽车，在我国的南方则可以提供小巧、精致的汽车。在技术方面，需要考虑技术的先进性，特别是高科技行业，技术的更新换代很快，要特别注意技术风险。

企业发展战略的类型主要包括成本领先战略、差异化战略、集中化战略、一体化战略。

如果企业要实施成本领先战略，首先需要确定在整个价值链中可以降低成本的链条。这些环节可能是采购环节、生产环节、分销环节。在采购环节，企业可以利用大数据对全球多个供应商的采购成本进行对比，找出成本、质量等满足企业要求的产品和服务。如果在生产环节，企业可以通过流程优化，找出在生产过程中可以去掉的工序。如果在分销环节，企业可以通过构建互联网、自助终端、迷你终端、移动终端等电子渠道，降低分销成本，并借助 O2O 协同，发挥分销渠道的整体优势。

如果企业要实施差异化发展战略，可以体现在产品、渠道、服务、销售、价格等方面。比如为客户提供优于竞争对手的产品，具有竞争对手没有的产品特征；可以比客户覆盖范围更广的渠道服务体系，提供线上和线下相结合的渠道体系；对于服务方面，比如缩短客户通过实体营业厅、呼叫中心的等待时间，可以根据客户价值高低分配客户的等待序列，实现差异化服务；在销售方面，可以提高业务开通的高效率，让客户能够更快地使用业务。

如果企业要实施集中化战略，首先要理解集中化对企业发展带来的好处，通常人们会认为资源集中能够发挥规模经济优势，同时也有利于企业管理。以企业信息系统的集中化为例进行说明，由于历史原因，采用两级或者多级组织架构层次的企业在各个层级都建设了信息系统，随着网络技术和信息技术的发展，为企业建设集中化的信息系统提供了可能，而且建设集中化的信息系统可以降低企业总体建设和维护成本，借助信息系统的集中化，也可以达到规范企业流程，增强企业统一管控能力的目的。

企业一体化发展战略可以分为横向一体化、前向一体化、后向一体化几种类型。横向一体化主要是企业需要收购竞争对手，完善渠道体系，扩大市场份额；前向一体化主要是企业收购或兼并一些渠道商，提升渠道控制能力；后向一体化主要是指企业收购供应商，

更好地保障企业原材料、服务等的供应。当然，企业内部也可以实施一体化战略，打通企业前后台部门和横向部门的流程，实现业务财务一体化、资源资产一体化和服务营销一体化，提升企业的整体运营能力。

1.2.3 物质决定意识：基础设施生命周期管理

当企业制定了发展战略，形成了企业的愿景、蓝图、目标等后，下一步就需要根据企业发展战略，进行基础设施的构建了。

按照企业架构分层的方法，企业首先需要定义出市场营销能力需求和产品能力需求，然后定义服务能力需求，再提出对于资源的能力需求和供应商/合作伙伴的能力需求。下面通过一个场景说明从能力需求到能力供给的过程。

比如企业制定了一个 4G 发展战略，在基础设施生命周期管理过程中要解决以下问题：

第一，企业发展 4G，需要考虑在哪些区域、针对哪些客户群进行市场营销，企业的产品能力是什么。比如，因为 4G 业务最大的特点是移动上网速度快，但是 4G 网络消耗的成本高，可以考虑首先在北京、上海、广州这样的发达城市进行网络建设，针对的是月流量消费在 300MB 以上的客户群。在产品能力方面，4G 新产品中主要具备用户自主定制上网流量、语音、数据业务的能力，同时根据在网客户的历史消费推荐适合客户的产品。

第二，需要确定相应的服务应当具备的能力。产品主要是面向客户的，而服务则需要面向内部。一个产品可能由多个服务组成，比如 4G 产品由 4G 移动上网服务、4G 语音服务以及短信数据业务组成。服务能力依赖于资源能力和供应商/合作伙伴能力来实现，比如 4G 移动上网业务需要北京地区采用自建的方式从供应商获取网络资源支持，也可以采用租赁的方式获取合作伙伴的 4G 基站资源。

第三，需要确定相应的资源应当具备的能力。资源分为逻辑资源和物力资源。逻辑资源比如交换机上的虚拟端口号，物理资源如交换机、路由器等的物理端口。服务是逻辑上对资源的抽象定义，最终还是要映射到物理资源。比如 4G 业务需要 4G 铁塔（BTS）、4G 基站（BSC）、互联网出口网关（GGSN、SGSN）等资源来承载。

第三，需要确定相应的供应商/合作伙伴具备的能力。当然，对于一个电信运营商来说，不一定所有的资源服务都是自家的，考虑到成本、交付速度等因素，会从供应商和合作伙伴处获取产品和服务。比如 4G 数据增值服务如天气预报内容服务来自于气象台，气象台

作为内容提供商，为电信运营商提供气象信息，形成天气预报数据增值服务。当然，自建4G基站等移动上网服务则需要通过采购比如华为、中兴这样的设备供应商的产品，最终形成4G移动上网服务。

同任何生物的生命周期一样，基础设施同样要经历从需求定义、设计、采购、验收、上架、上线、下线、下电、报废等一系列过程。当由于技术过时、设备老化等因素导致市场不再需要这个基础设施时，经过业务影响评估等过程，基础设施需要完成下线下电报废等处理，退出服务的提供。

1.2.4 你我约定：产品生命周期管理

犹如所有生物都会经历从出生、成长、衰退直至死亡的过程一样，企业的产品也同样会经历从创意、设计、开发、上线到下架的一系列过程。企业内部典型的产品上线过程如图1-2-3所示。

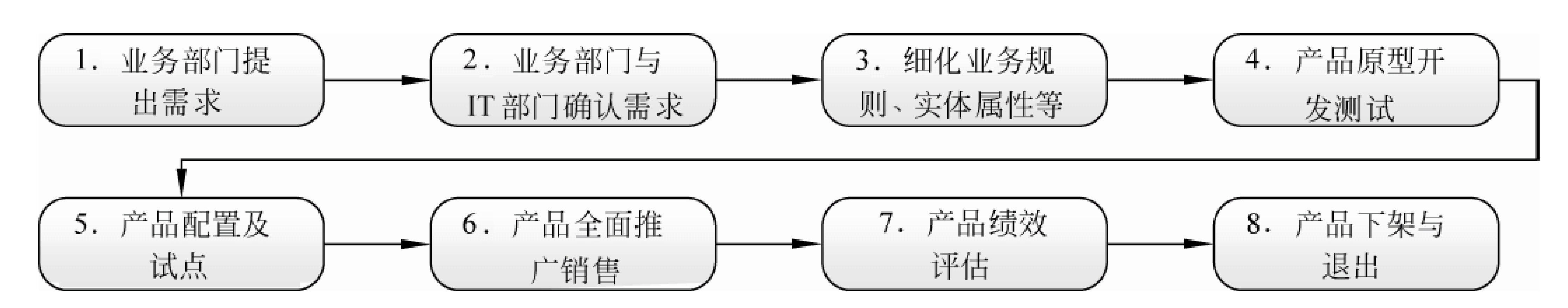


图1-2-3　产品生命周期过程示例

产品是一个企业价值转换的媒介物。产品分为有形的和无形的两种类型，有形的产品如手机、SIM卡、手机配件等，无形的产品称为服务，比如通信套餐、金融理财产品等。

按照市场/产品/客户、服务、资源、供应商/合作伙伴的分层方法，在市场/产品/客户层，产品生命周期管理细分为产品提供品（offer）开发与管理、产品营销传播与推广、销售开发三个阶段，服务层包括服务开发与管理过程，资源层包括资源开发与管理过程，供应商/合作伙伴层包括供应商/合作伙伴开发与变更管理过程。下面对三个阶段和不同支撑层次的过程进行解释。

第一，产品提供品开发与退出、产品营销传播与推广、销售开发：当基础设施能力具备后，企业就可以设计满足市场需求的产品了。产品开发包括从产品创意、产品设计、审批、研发、实验等一系列子过程。产品开发完成后，还需要针对市场和客户群实施营销宣

传，包括营销方案制定、产品宣传等。销售开发过程包括制定产品补偿的政策、开发新的销售渠道、销售培训等，销售开发通常以项目为管理单位，而运营过程中的销售是日常型的（day to day）。

第二，服务开发与退出：产品开发过程定义了产品的规格、价格、销售渠道等要素，但是要想让客户享受到产品还需要进一步定义服务。产品更多地关心客户的需求、企业对于满足这些需求的价格回报以及推广客户与产品的渠道等，而服务更关心如何将价值交付给客户。比如电信运营商推出一个“亲情一家产品”，这个产品由移动、固话和宽带三个服务（也称为业务）组成，这些业务的价格分别为：移动业务每分钟四角，固话费用为每分钟两角，宽带月租费为 100 元，“亲情一家产品”通过实体营业厅销售。

第三，资源开发与退出：服务是从产品分解来的，但是服务只是一种能力的抽象，服务需要资源的支持才能够真正地为用户提供服务。按照资源的表现形式，将资源分为逻辑资源和物理资源两类。逻辑资源是虚拟的，比如 IP 地址、电话号码、端口号等。物理资源则是有形的，看得见摸得着的，比如主机服务器、存储设备、网线、网卡上的端口等。

根据服务的要求，需要进行资源的开发，比如宽带接入服务，它需要人工和自动服务共同完成，对于人工服务需要宽带猫、网线、工具等物理资源，对于自动服务，需要号码、用户等逻辑资源。

第四，供应商/合作伙伴开发与变更：企业的资源可以采取购买或租用的方式从供应商/合作伙伴获取，因此要建立与供应商/合作伙伴合约、承诺、协议的签署与变更，保证企业能够按照约定来提供资源和服务，否则企业可以变更供应商/合作伙伴。采购周期通常为：用户需求—>确定规格—>决定生产还是购买—>竞标还是议标—>供应商选择—>供应商关系管理—>用户需求。

1.2.5　各就各位：运营支持与就绪过程

企业在发展战略的指导下，实施了基础设施生命周期管理和产品生命周期管理，完成了从战略、建设到产品供给的过程，为企业运营做好了铺垫。

但是，企业要实现正常运营，还需要相关的辅助支持过程，以保证企业运营的顺利进行，将这个过程称为运营支持与就绪（运营准备）。

（1）客户接触管理支持过程：核实客户接触管理过程是否已经具备相应的能力。客户

与企业可能通过多种渠道接触，包括电话、网站、代理商、短信等。比如企业已经具备了4G产品销售能力，为了支持客户通过电话、网站渠道进行产品咨询，客户接触管理支持过程应当核实IVR脚本是否定义，4G产品宣传所有的录音文件是否已经准备好，是否已经将4G产品信息嵌入网页中等。如果还不具备这些能力或者存在问题，则应当及时进行修正，以便客户接触管理过程的顺利进行。

（2）市场营销实施支持过程。

（3）销售支持过程。

（4）订单处理过程支持过程。

（5）问题处理过程支持过程。

（6）账单查询处理过程支持。

1.2.6 不仅是讨价还价：售前阶段的业务过程

运营支持过程完成了运营前的准备工作，就好像飞机起飞前检查油箱是否有足够的油、起落架是否正常一样，目的是为了保证飞机的正常飞行。当检查完毕并解决好存在的问题后，企业就可以正式运营了。

从面向客户销售的角度看，企业运营可以划分为售前、售中、售后三个阶段，其中售前阶段主要活动是企业与客户之间通过反复沟通，就价格、方案等达成一致意见。售中阶段的主要活动是按照售前阶段双方的约定，完成服务的开通，使得客户能够获得企业提供的产品或者正常使用企业提供的服务。

企业售前阶段的业务过程如图1-2-4所示。

从图1-2-4可以看出，首先在企业市场宣传推广的驱动下，客户接收到企业产品信息，然后通过各种渠道（网站、代理商、零售商等）与企业建立联系。其次是客户将需求提供给企业，企业根据客户需求进行方案设计，方案设计阶段企业需要核实自己是否具有满足客户需求的资源和能力，如果暂时不具备，可以通过从供应商/合作伙伴采购的方式获得生产用资源。当确定解决方案后，企业再为客户提供不同的销售建议，通过双方沟通确认，形成销售意向。

对于服务的开通过程来说，客户总是希望尽可能少地等待就可以使用企业提供的产品和服务，因此企业应当尽快交付产品和服务，以便提升客户满意度。

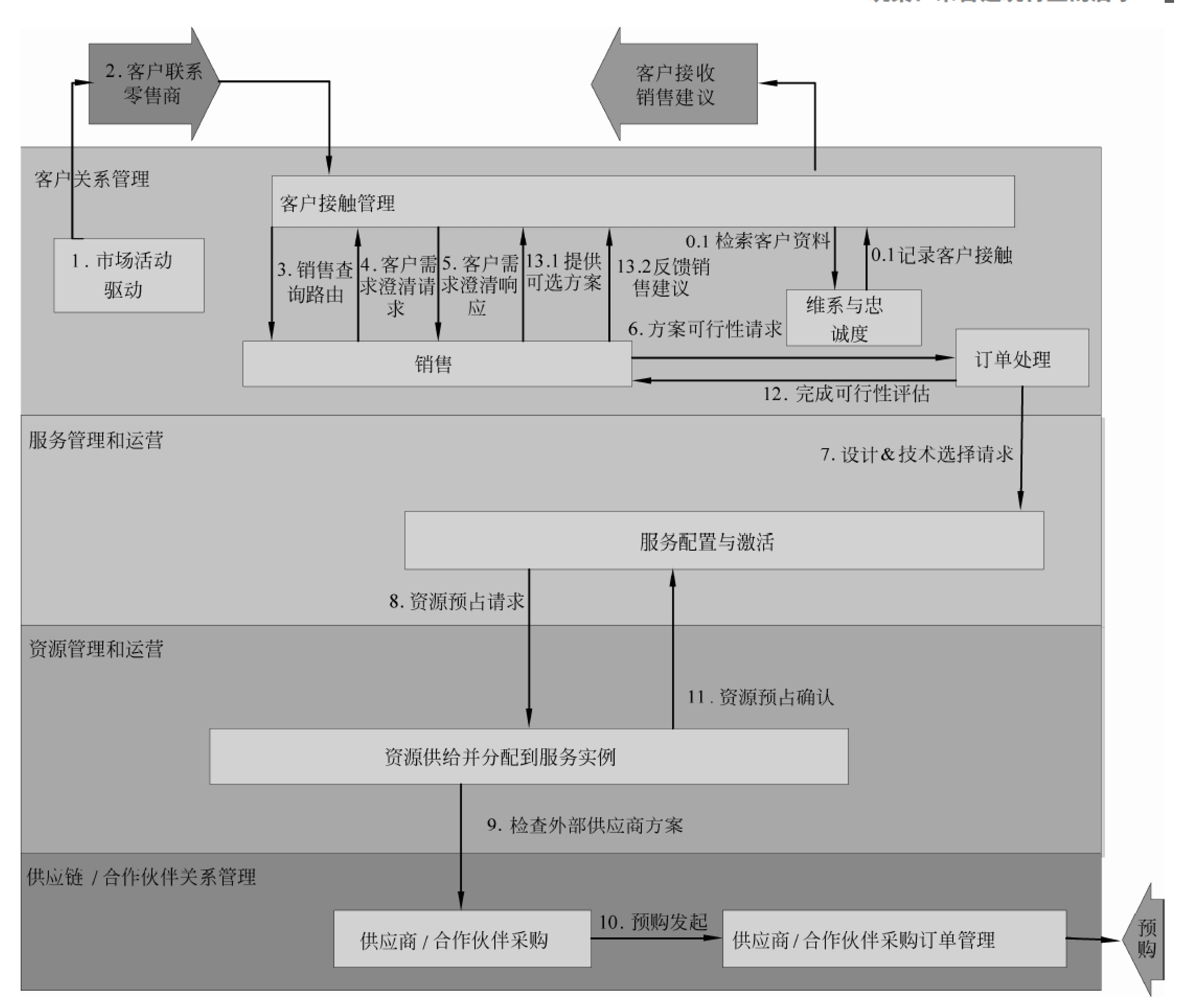


图 1-2-4　企业售前阶段的业务过程

1.2.7　零等待靠谱吗？售中阶段的业务过程

在售前阶段，企业和客户之间达成意向并“预占”了资源。当意向变成协议后，企业就需要为客户“开通”服务了，由于这个阶段企业还没有完成产品或者服务的销售，因此称之为“售中”阶段。

售中阶段其实是一个客户需求落地实施的过程，企业根据与客户预先的约定和方案，将业务订单分解成多个执行工单，当各个工单全部实施完成并报竣后，意味着客户从此能够使用企业的产品和服务了。

企业售中阶段的业务过程如图 1-2-5 所示。

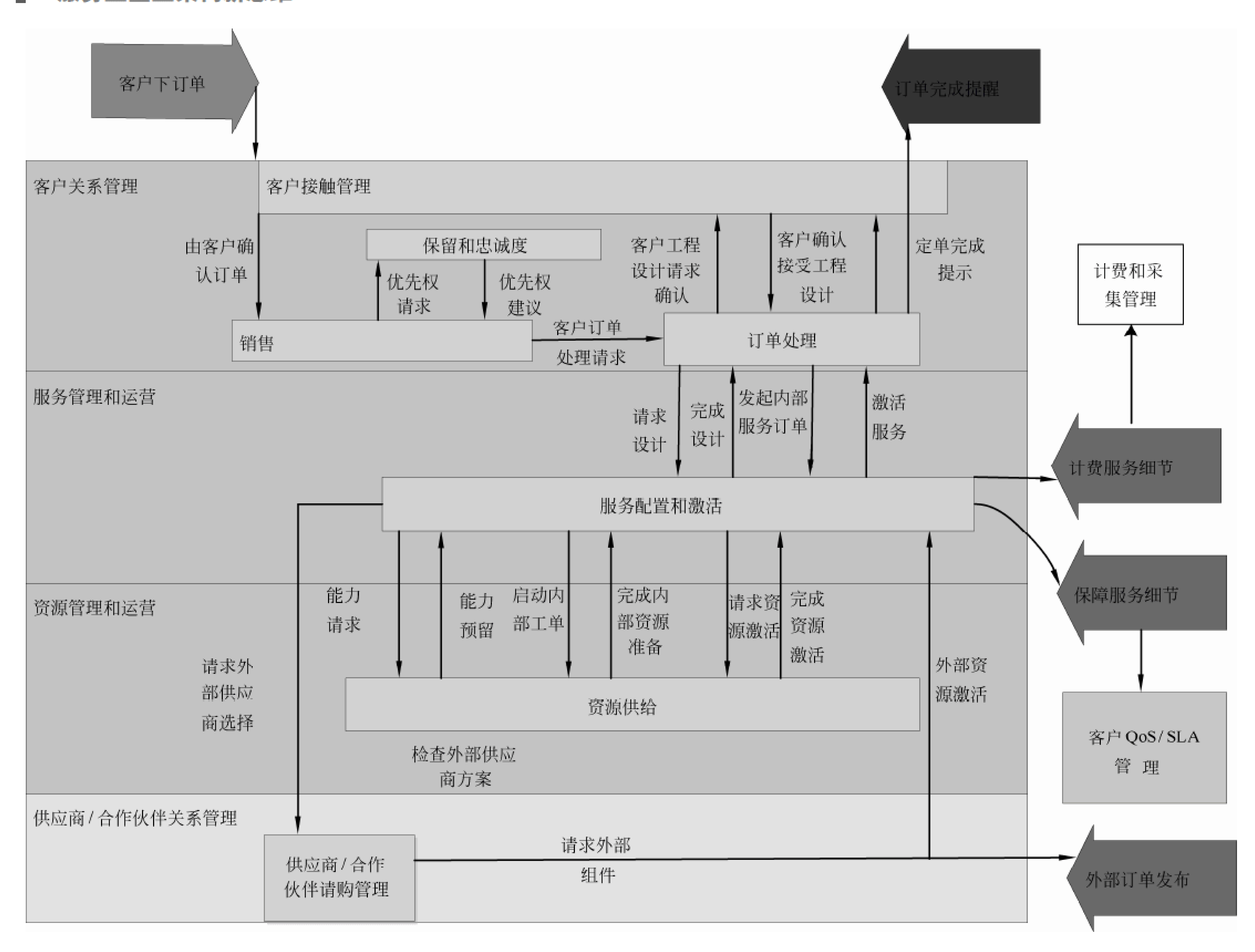


图 1-2-5　企业售中阶段业务过程

从图 1-2-5 可以看出，售中阶段实际上是从需求（订单）到供给（工单）的实施过程。订单表示已经“确定”，体现了企业与客户之间就需求达成的一致意见，比如产品的规格、价格、售后服务、交付时间、交付地点、交付方式等。企业会以订单为输入，按照流程和规则，将订单分解为多个可以执行的工单（操作单）。工单实施分为自动和人工两种方式，比如服务号码的预占、实占和释放，就是通过改变号码状态在系统中自动完成的，比如宽带和固话的安装，需要安装工人根据安装地点、配置、时限要求等在现场完成配线、配号、配端口等工作。企业应当尽量采用自动实施工单或者客户自助服务的方式来降低总体成本并降低因为人工操作导致的错误，对于复杂的操作才考虑采用人工方式。

1.2.8　前后台的双簧：售后阶段的业务过程

企业开通为客户提供的服务后，客户就可以使用企业的产品了。客户在使用产品的过

程中，可能会出现故障，比如电话掉线、无信号、无法上网、多收费等问题或者对于产品的资费、服务网点等咨询问题。

那么，如何解决来自于客户反馈或者企业内部发现的问题，成为服务保障过程考虑的内容。售后阶段的业务过程如图 1-2-6 所示。

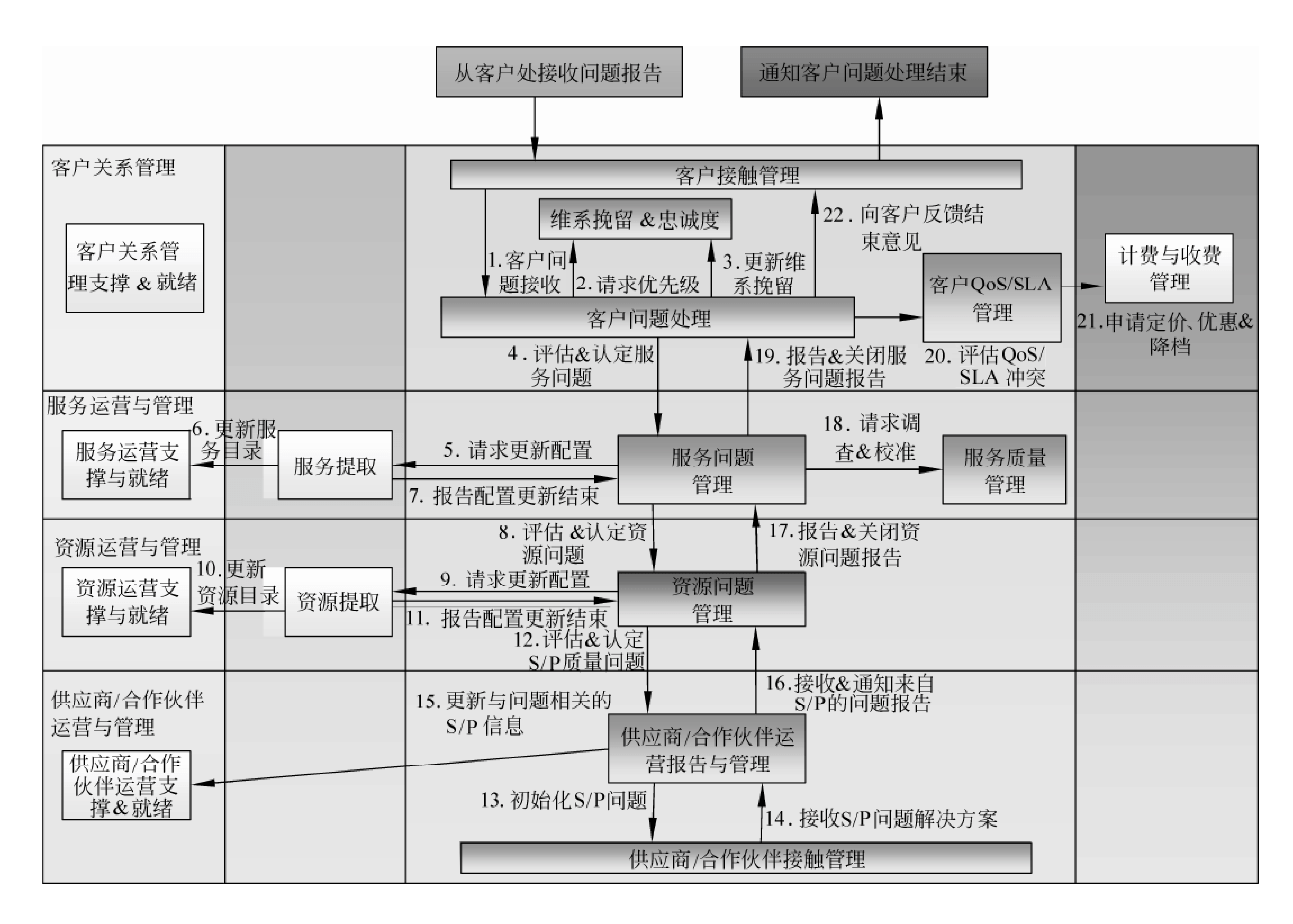


图 1-2-6　企业售后阶段业务过程

从图 1-2-6 可以看出，企业从客户接收到问题并对问题进行处理，比如由企业客户服务中心的话务员对问题进行初步分类和记录，然后分发给服务问题管理过程，服务问题管理过程对问题进行诊断，更新服务目录，再将问题发给资源问题管理过程，资源问题管理过程进行诊断，更新资源目录，然后将问题发给供应商/合作伙伴问题管理过程，该过程同样是更新供应商/合作伙伴信息，然后将问题发给供应商/合作伙伴处理，等接收到处理结果后，将各个过程环境的问题处理结果汇合起来，然后将最终处理结果反馈给客户。

1.2.9 无利不起早：企业计费收费过程

企业通过为客户提供价值来获取价值，这是一种交换。当客户购买产品和服务时，订单中体现了双方价值交换的内容，约定了产品和服务的“价”，而客户使用产品和服务的“量”，则是在客户使用企业提供的产品和服务的过程中形成的。下面就通过电信运营商、银行、互联网公司简单看一下计费收费的过程。

对于电信运营商来说，从计费到收费/交费包括采集、批价、计费、出账、交费充值几个步骤，计费的数据基础为业务使用记录，暂时称之为 xDR（通话记录、上网记录、短信次数等）。在 xDR 采集阶段，需要收集话单、上网记录、增值业务使用记录等。采集的 xDR 经过去重、格式化等处理后形成批价的输入数据源，然后计费系统再根据批价规则进行批价，根据优惠规则进行计费，然后将各种业务的消费情况进行汇总形成账单后出账。对于用户，可以通过营业厅、网站、充值电话等渠道进行充值缴费。

不同于电信运营商具有庞大的网络资源，以银行为代表的金融企业的主要收入为贷款人的利息和服务费用。银行主要通过对于资金的有效配置，收集社会闲散资金并将资源配置到那些需要资金的环节中。与电信运营商类比，银行的供应商为存款单位和个人，客户是从银行取得贷款的企业和个人，银行的职能是对这些资金进行运营管理，此外，银行也受银监会等监管机构的监管，存贷款需要在一定的规则下完成。为了扩大对资金的利用，在以存贷款业务为主的银行之外又衍生出了多个金融机构，比如保险、证券、金融租赁等，这些机构的经营模式与银行又有很大的不同。

对于互联网公司来说，商业模式有很大的不同。以提供信息服务的门户网站为例，它们的商业模式主要是为公众客户提供信息服务，往往是免费的，但是互联网公司可以在门户网站上植入广告，向发布广告的机构收取费用，业界称之为反向收费模式；对于电子商务公司，包括平台模式、自建自营、自建他营等诸多商业模式，对于平台模式，互联网运营商通常向入驻商家收取交易佣金，对于自建自营模式的互联网运营商，与实体店类似，目标是赚取产品的销售利润。

1.2.10 无声的发动机：企业内部管理业务过程

企业对外完成市场经营活动，实现产品与服务的营销、销售及服务工作，这些工作为

企业带来收入和利润，但这些前台光鲜的工作缺少不了企业后台的管理做支撑。

企业的内部管理业务过程主要包括人力资源管理、财务管理、资产管理、工程项目管理。此外，还包括协同管理、风险管理、知识管理等。

“人”是服务型企业管理的关键，人力资源管理包括招募、用工、选拔、薪酬、绩效、离职、转岗、合同等业务活动。“财”是服务型企业的管理核心，包括财务预算、会计核算、资金支付、资金稽核、财务报账等业务活动。“物”是服务型企业管理的重心，包括资产录入、资产盘点、资产折旧、资产报废等业务活动。下面分别介绍一下企业内部各个过程的内容。

1. 人力资源管理过程

人力资源的管理对象是人，人力资源管理以人为管理中心，包括人员基本信息管理、招聘管理、薪酬管理、绩效考核管理、培训管理、职业规划管理、考勤管理等几个方面，涵盖员工招募、团队建设、激励、培养、退出的全生命周期管理。

人员基本信息管理包括员工编号、姓名、学历、工作经历、教育经历、岗位、职称等信息；招聘管理包括人员需求管理、招聘渠道管理、招聘信息发布、简历筛选、招聘通知、笔试管理、初试管理、复试管理、录用管理等；薪酬管理包括基本工资管理、工资结构管理，工资结构中包括具体的工资项，如基本工资、书报费、取暖费、洗理费、社保、公积金等；绩效考核管理包括考核指标管理、考核报表管理、考核分析；培训管理包括培训机构管理、培训讲师管理、培训需求管理、培训效果评估等；职业规划管理包括职业路线管理（比如技术路线、管理路线）、职业规划访谈、职业规划指导、职业规划推荐等；考勤管理包括签到管理、签退管理、考勤统计、事假管理、病假管理等。

2. 财务管理过程

应收（AR）、应付（AP）、固定资产（FA）、总账（GL）是财务管理的核心内容。采购（PO）、库存（INV）、项目会计（PA）、项目开单（PB）、现金管理（CE）等过程也与财务管理过程有着密切的关系。

应收，即企业应当收取客户、合作伙伴的费用；应付，即企业应当支付给内部员工、供应商、合作伙伴的费用；固定资产，即办公、网络、维修工具等方面的企业内部资产，记录了资产的原值、折旧率、折旧后的费用等；库存管理，即管理定制终端、网络设备、管理工具、办公用品等暂时存在仓库待使用的物品。总账管理是财务管理的核心过程，主

要完成会计和财务信息的记录和集取，完成财务监测和控制以及财务信息的分析和报告。

1.2.11 本节小结

从企业创办之日起，就会与企业外部的客户、供应商、合作伙伴以及企业内部的股东、雇员等利益相关者交互，开展战略、建设、运营、企业管理等业务活动，完成价值的创造与交付。

由于企业战略、建设、运营、管理的业务活动相互联系、相互制约，为了更好地对业务活动进行管理，需要采用系统化的思维，采用分层分类的方法，实现对企业业务活动进行有效的管理。

本节从时间轴和空间轴两个维度，将企业业务活动划分为矩阵型的过程块，过程块之间相互配合，协同完成企业面向战略、业务、管理方面的全部职能。

时间轴角度划分过程块是过程化思维，从总体上将企业过程切割成从企业战略、基础设施生命周期管理、产品生命周期管理、运营支持与就绪、售前、售中、售后、计费收费、企业管理等多个子过程。

空间轴角度划分过程块是结构化思维，以客户为中心，按照从客户需求到客户供给的思路，由外到内将企业支撑结构分为市场、产品、客户、服务、资源、供应商/合作伙伴几个层次。

总之，从时间维和空间维两个视角，对企业业务过程进行管理，形成企业业务过程管理的整体框架，按照分层的方法，将企业业务过程划分为不同层次的过程块。

1.3 信息：企业业务活动的承载者

信息与业务过程是一体的、不可分割的，业务过程是动态的，信息是静态的，两者相互配合，组成了各种各样的业务活动。

人与人之间在工作与生活中需要交流，交流的内容会以信息为载体传达给对方。在组织中，信息是在业务过程中形成的，因此，作为管理信息的信息框架与管理业务过程的业务过程框架是一体的、不可分割的。

如果说业务过程是动态的，可以用动词来定义，那么信息则是静态的，可以用名词来定义。为了理清企业内部各种信息之间的关系，首先需要确定参与信息交互的业务对象，然后再确定业务对象之间的关系。业务对象也可以称为参与方或者实体，与业务过程按照层次划分不同，业务对象按照颗粒度进行划分。业务对象之间的具有不同的关系，比如包含、继承等。

1.3.1 概念模型

与数据相比，信息是属于需求域的。信息模型是对需求的刻画，描述了需求域中业务对象之间的关系，通常也称为概念模型，如图 1-3-1 所示。

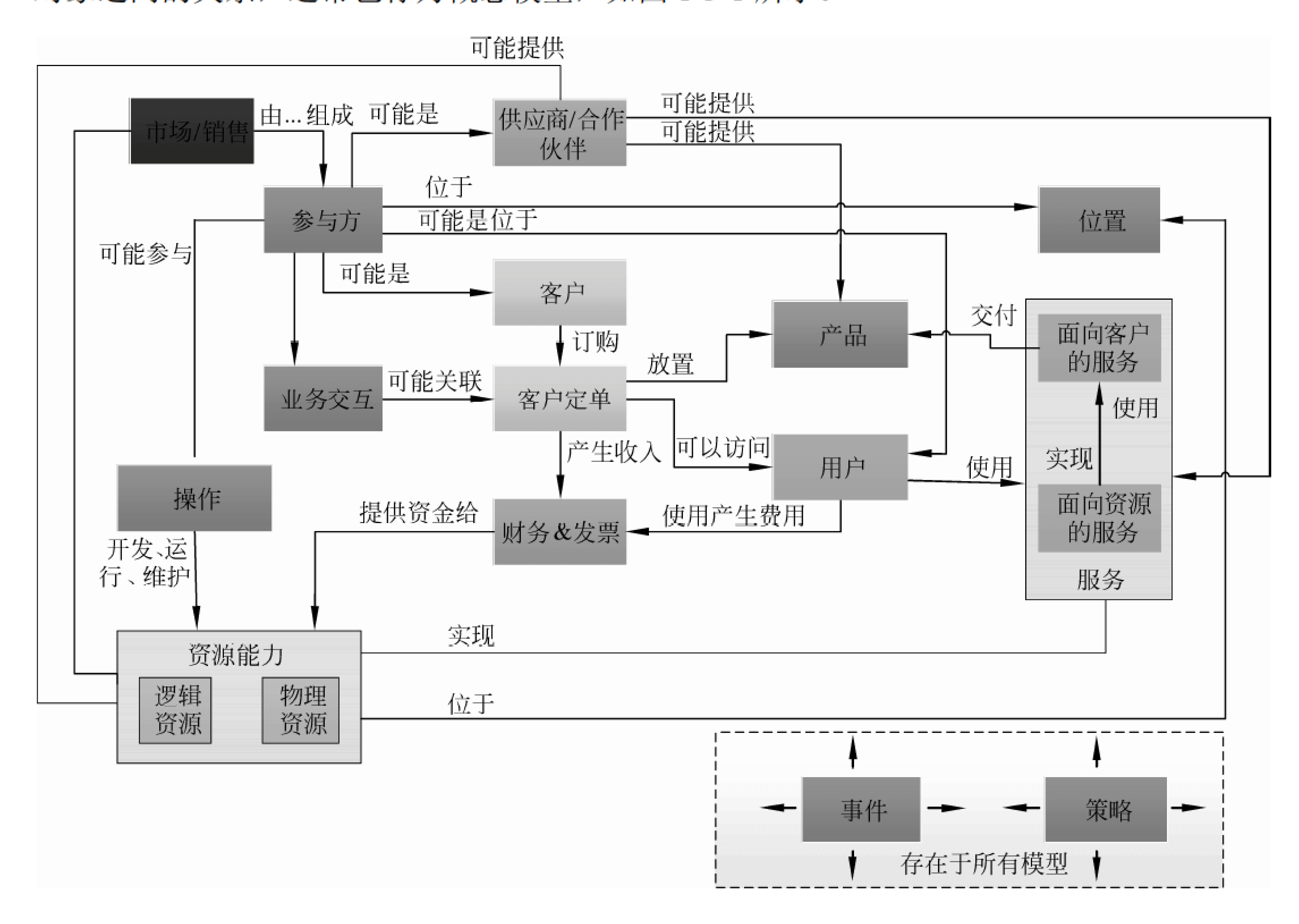


图 1-3-1　企业概念模型示例

之所以将信息模型称为概念模型，是因为它是对现实世界的一种表达。企业概念模型表达了企业内部各个业务对象之间的关系，通过概念模型，可以直观地看到现实世界中各

个对象以及它们之间的联系。

从企业运营的角度看，一个企业首先应当具有自己的产品，当然必然会有自己的客户，那么就会有“客户”和“产品”这两个概念模型。

其次，从企业为客户提供产品的视角看，企业会通过“市场/销售”的手段将“产品”传递给“客户”，如果“客户”与企业之间达成共识，那么“客户”就会通过“客户定单”的方式来实现自身需求，当“客户定单”完成后，对于“企业”则需要通过“财务”计入应收账款，对于“客户”则在支付产品或服务的费用的同时应当得到企业提供的“发票”。当企业完成以上活动后，“客户”就可以使用企业提供的产品了，这时候的“客户”就变成“用户”了。

再次，从企业“产品”形成的过程看，形成“产品”的物质基础是“资源”，而“资源”也分为相互联系的两种类型：物理资源和逻辑资源。“物理资源”是看得见摸得着的“资源”，比如一台服务器的硬件设备，包括机箱、主板、CPU、内存、硬盘、网卡、显卡等。而“逻辑资源”则正好相反，它是人类通过人脑抽象出来的，是看不见也摸不着的，比如人们常常说的手机号码、IP 地址、逻辑端口号等。逻辑资源是人类为了便于管理而设计的，它就像给一个人取一个名字，以便与他人区分，抽象是人类特有的。因此“逻辑资源”比“物理资源”要灵活，就好比哲学中抽象和具体的关系，同时“逻辑资源”必须有“物理资源”作为物质基础，比如在实际应用中，IP 地址必然对应着一台具体的机器设备。

然而，无论是“物理资源”还是“逻辑资源”，在面向外部市场方面，都存在着不足，原因是客户对于产品的需求通常不是单一的，对于企业来说，为了市场营销和销售的需要，也常常会对多个“产品”进行打包。为了解决这种供需之间的矛盾，引入了“服务”概念模型。“服务”不同于“资源”，为了面向市场中客户的差异化、多变的需求，“服务”可以对多个“资源”的能力进行组合后形成“产品”的基本结构，然后再以此为基础，增加面向市场的其他元素，比如市场细分、渠道、价格、SLA 等，最终形成一个面向市场的完备“产品”。

当然，“服务”在将处于“供给”侧的“资源”转变为面向“需求”侧的“产品”的过程中，也不是一下子就完成的，也需要一个从“面向资源的服务”到“面向客户的服务”的转变过程，之所以有这样的划分，主要是消除由于“供给”和“需求”的关注点不同而引发的问题，实现平滑过渡。

最后，虽然解决了从企业内部“资源”供给到企业对外“产品”提供的转变，但是“客户”、“用户”、“供应商”、“合作伙伴”等业务对象之间如果建立连接关系，以上概念模型

将变得非常复杂难懂，为了将业务对象之间的复杂关系简单化，增强概念模型的灵活性，引入了“参与方”。

此外，像“事件”、“策略”属于共享型的概念模型，因此不单独设计，以降低概念模型的复杂性。

以上为企业的整体概念模型，概念模型还可以细分，通信设备的概念模型如图 1-3-2 所示。

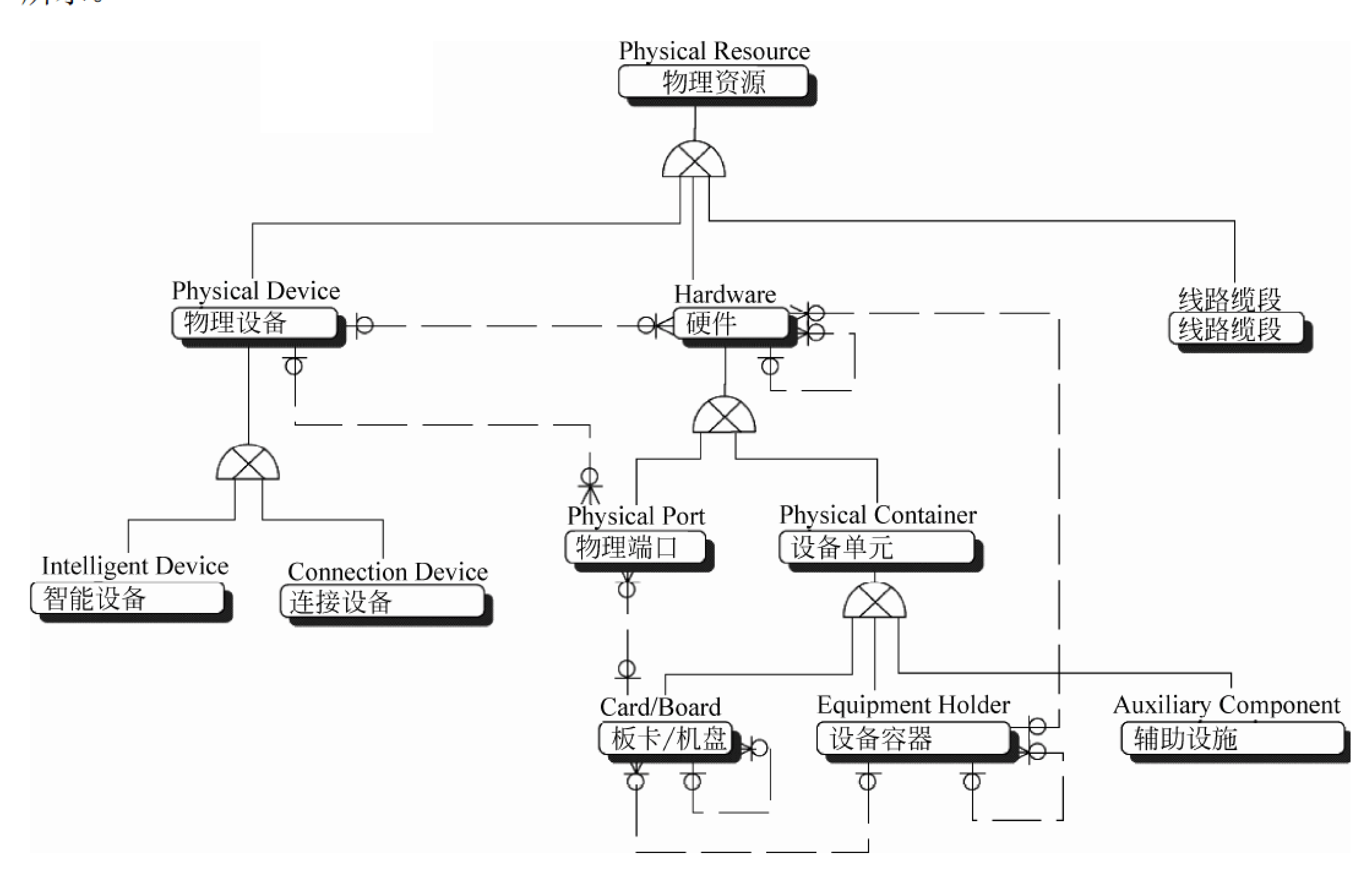


图 1-3-2　通信设备概念模型

从图 1-3-2 可以看出，由于概念模型与现实世界比较接近，因而比较容易理解。物理资源包含物理设备、硬件和线路缆段，硬件包括设备单元和物理端口，通过概念模型构建以上设备之间的逻辑关系，可以将现实中的设备在软件系统中有效地管理起来。

1.3.2　信息框架

概念模型虽然可以从业务视角对需求进行刻画，但是如果概念模型增多，将变得难以

管理，为了实现对概念模型的有效管理，引入了信息框架。信息框架与业务过程框架的管理方式类似，同样采用分域、分层的方式进行管理。信息框架与业务过程框架相对，同样是分为市场/销售、产品、客户、服务、资源、供应商/合作伙伴、企业，共 7 个域，此外，还有一个特殊的公共业务实体，比如参与方、项目、位置、协议等。信息框架的一级结构如图 1-3-3 所示。



图 1-3-3　信息框架示例（一级）

为了直观地看到信息框架和业务过程框架在业务需求管理中的一体两翼关系，下面对这两个框架进行对比，如图 1-3-4 所示。

从图 1-3-4 可以看出，业务过程框架中的第一层（市场、产品、客户）在信息框架中被分为市场/销售、产品、客户三个独立的域。其他域如服务、资源、供应商/合作伙伴、企业管理则表现为一对一的关系。此外，与业务过程框架不同，为了体现业务对象之间的

复用性，降低业务对象之间的复杂性，新增了一个公共业务实体域。

图 1-3-4　业务过程框架与信息框架对比

1.4　应用：业务与技术之桥

应用即能力，它填平了业务与技术之间的鸿沟，是业务与技术之桥，应用框架又称为能力蓝图，体现了业务人员与技术人员的共同愿景。

业务过程以及业务过程中产生的信息是对业务需求的描述，但是要想提高企业的运营效率和竞争能力，还必须借助信息技术，将业务需求落地到信息系统之中。为了理清业务需求与信息系统之间的“界面”，引入了“应用”的概念。

从业务和技术的特点看，业务是对现实活动的抽象，业务过程刻画了企业为完成自身使命需要做的“事情”，比如企业要为客户提供产品，必须从供应商处采购到原材料，从市场上招募到合适的人员，取得企业经营所需的资金等，企业的这些活动与是否存在信息技术是没有什么直接关系的。

但是，自从信息技术的出现，企业就不能按照古老的方式经营了，原因是信息技术可以提高企业的生产和经营效率，提高企业的经营管理能力。因此，企业要想借助信息系统支撑各种业务活动，首先需要将业务需求转化为信息技术能够接受的需求。然而，业务与

技术之间存在天然的鸿沟。比较而言，业务语言更加贴近自然语言，相对发散，而技术语言是机器语言，一就是一，二就是二，相对收敛，这也就形成了业务人员和技术人员在思维方式上的不同，为了让业务需求落地到信息技术实现，必须在业务语言和技术语言之间找到一个“中介”，而这个中介物就是“应用”。

“应用”也可以叫作“能力”，对于业务人员来说，他们不一定要掌握各种信息技术，但是他们可以提出对于信息技术的“能力”需求。同样，对于技术人员来说，他们不一定精通业务，但是可以按照业务人员的“能力”需求完成信息系统的设计与实现。这样，业务人员和技术人员就能够以“应用”为纽带，实现沟通和理解。比如，业务人员可以对信息系统提出业务查询响应时间在 3 秒以内的“能力”需求，那么技术人员设计的信息系统应当能够在 3 秒之内反馈查询结果，这就是业务人员和技术人员对信息系统达成的“能力”共识。

“应用”除了承担业务人员和信息技术人员之间“共同语言”的角色之外，在买方和卖方之间还充当“合约”的角色，买卖双方可以“应用”作为标的物，确定产品的能力要求、价格等要素。

1.4.1　应用框架/能力蓝图

应用是相关功能以及其他相关应用的集合体。应用在业务人员和技术人员之间搭起一座桥梁，实现双方的“理解”。业务人员可以对技术人员说：“你们需要实现这些能力，这是我们的需求，有了这些能力，我们的业务能力就强大了！”；技术人员也担心自己说的话太“技术”，业务人员听不懂，并且担心因为没有沟通好而白做了工作，于是问业务人员：“系统实现这些能力就满足你们的需求了吗？”业务人员明确地回答：“是这样的！”

应用框架也称为能力蓝图，可见应用框架是能力的集合体。应用框架是业务过程框架向技术实现的进一步收敛，同时也包括了公用的应用，那是因为应用具备了技术特征，而技术是可以复用的。

应用框架与业务过程框架相对应，从纵向看，包括战略、基础设施生命周期管理、产品生命周期管理、运营准备、服务开通、服务保障、服务计费，这些与业务过程的分类是一致的。从横向看，包括市场/销售域、产品管理域、客户管理域、服务管理域、资源管理域、供应商/合作伙伴管理域以及企业管理域，这些与业务过程框架基本上是一致的。此外，还包括交叉域和集成架构域，前者是其他域共用的，而后者则是实现应用之间集成而需要

的能力。一级应用架构如图 1-4-1 所示。

以产品管理域中产品生命周期管理应用为例，其能力要求如下：

- 提取产品需求。
- 产品建模。
- 提供详细的产品规格。
- 新产品引入。
- 管理现有产品。
- 产品废弃/退出。
- 市场与定价战略实施。

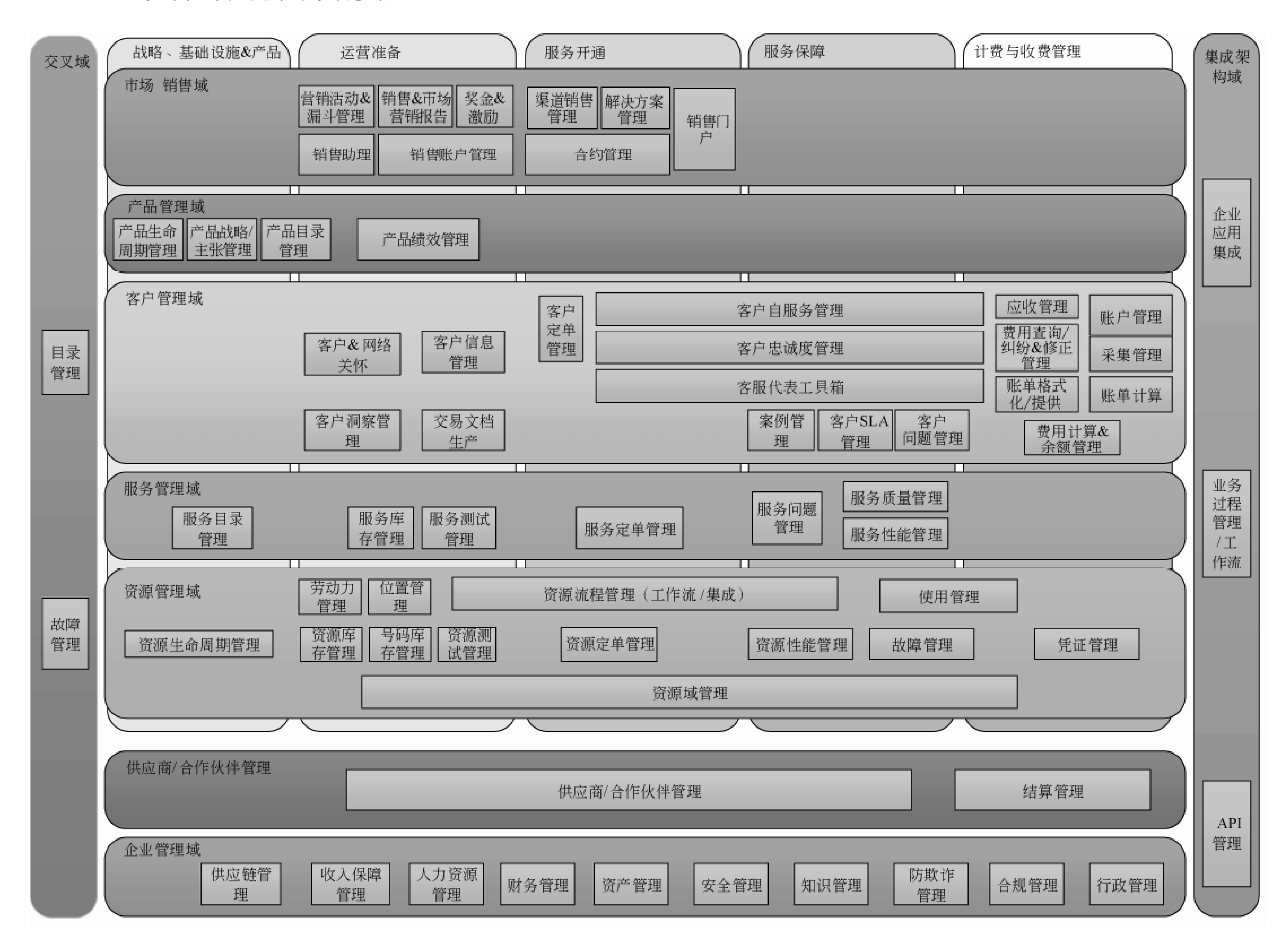


图 1-4-1　应用框架（1 级）

能力蓝图是企业对于目标或者当前信息系统具备能力的期望或者评价，通过能力蓝图，可以直观地看到企业对于信息系统能力的要求，便于企业信息系统能力差距对比分析，

也便于企业进行信息系统产品和服务的采购。

1.4.2 应用框架与业务框架

应用是业务向信息系统的收敛，因此应用必然来源于业务，为了清晰、直观地看到业务与应用的关系，下面就对其进行初步的对比分析，业务过程框架、信息框架与应用框架的总体对比如图 1-4-2 所示。

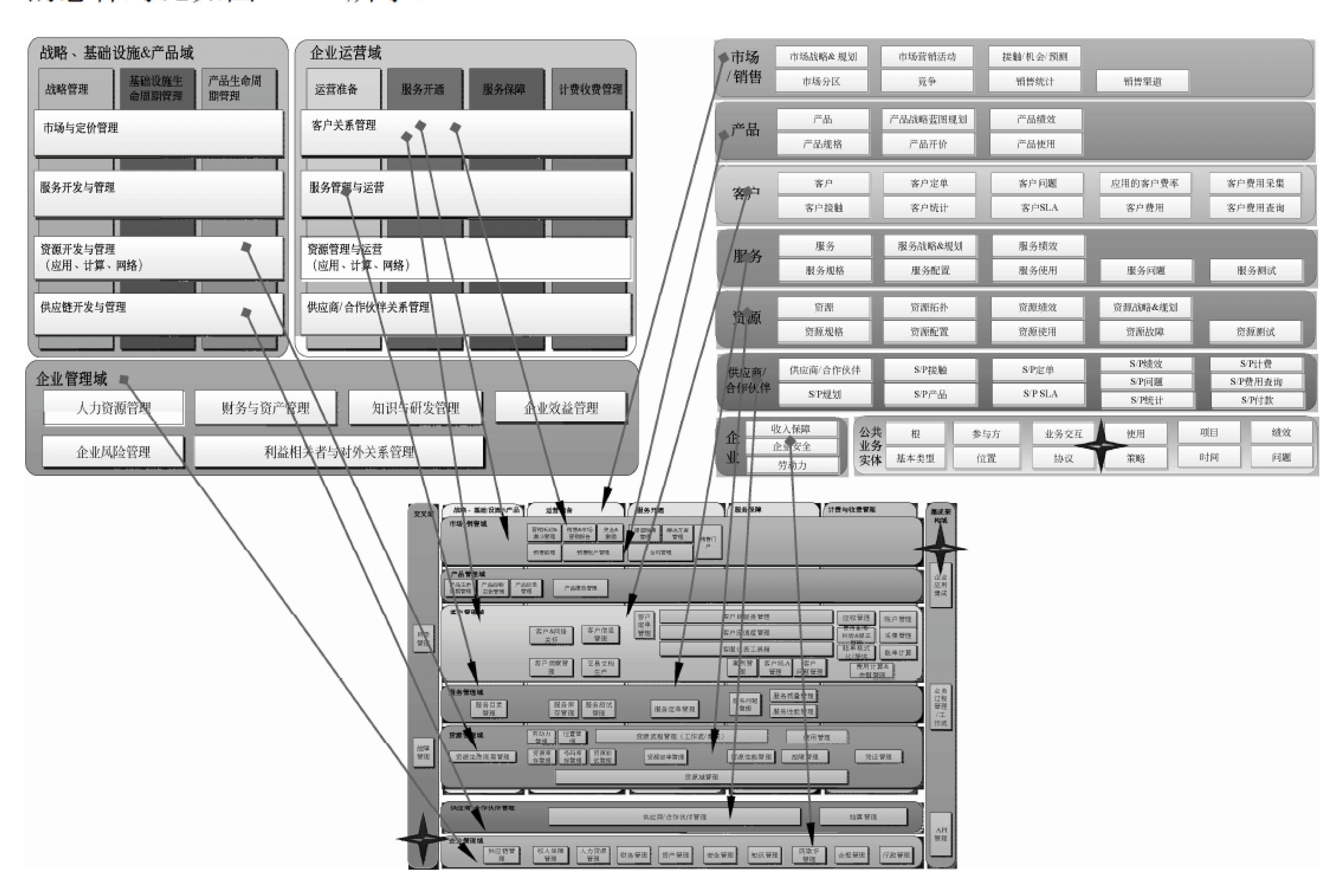


图 1-4-2 业务过程框架、信息框架与应用框架的关系

为了对应用进行有效的管理，应用框架内部划分为多个域。从图 1-4-2 可以看出，业务过程框架与应用框架之间既有区别又有联系，相同点是：基本上按照市场、产品、客户、服务、资源、供应商/合作伙伴、企业管理的分类方式进行管理，不同之处是应用框架不像业务过程框架那样分为企业战略、基础设施生命周期管理、产品生命周期管理以及企业运营两大部分，而是以市场/销售、产品、客户、服务、资源、供应商/合作伙伴为主题进行了收敛，其目的是消除业务过程框架中按时间段划分业务对象而增加的管理难度，这也是

“能力”视角与业务视角最大的区别。“业务”视角着重依照业务的开展情况来刻画业务本原，而“能力”视角则是将“发散”的业务需求“收敛”为多个不同的主题域，这样更便于信息系统的实现。

同理，由于对于信息的管理不必太关注业务实施的时间先后顺序，因此信息框架与应用框架类似，也采用面向不同主题的管理方式，以被管对象为中心进行了收敛。

应用框架与业务过程框架和信息框架除了以上的区别之外，还考虑了应用的复用问题，因此新增了公用应用支撑域，这个域内的应用是可以被其他域的应用所使用的，比如目录管理、故障管理等，这点与信息框架中的公共业务实体类似。

此外，应用框架中还新增了公共基础设施支撑域，公共基础设施支撑域同样是从复用性角度考虑的，包括服务总线、业务流程管理、中间件等。以业务流程管理为例，服务定单管理应用可以基于它来完成定单到工单的分解，产品生命周期管理应用也可以基于它完成产品从就绪到上架过程的转变，因此业务流程管理属于共享型应用。

1.5 功能：特定任务的执行单元

> 功能以应用/能力需求为输入，采用信息技术手段，将能力需求转化为用户可以使用的、具有特定规格要求的单元。

功能用于执行特定任务，是能力的实现。功能对应的英文名称为 function，维基百科对 function 的解释为：subroutine, a portion of code within a larger program, performs a specific task，中文意思为：功能是子例程，是大型程序中执行特定任务的一部分代码。

英国商务部 OGC 制定的 ITIL 第三个版本中对于功能有更加详细的定义：Functions are units of organizations specialized to perform certain types of work and be responsible for specific outcomes，中文意思为：功能是组织中专门执行特定类型工作的单元并且负责输出具体的结果。

可见，功能是与特定的任务/工作挂钩的，并且具有特定的输入和输出。功能框架就是功能的集合体，功能以应用/能力需求为输入，采用信息技术手段，将能力需求转化为用户可以使用的、具有特定规格要求的单元（unit）。

下面就以日常工作中经常接触的办公自动化系统为例对功能进行说明，如图 1-5-1

所示。

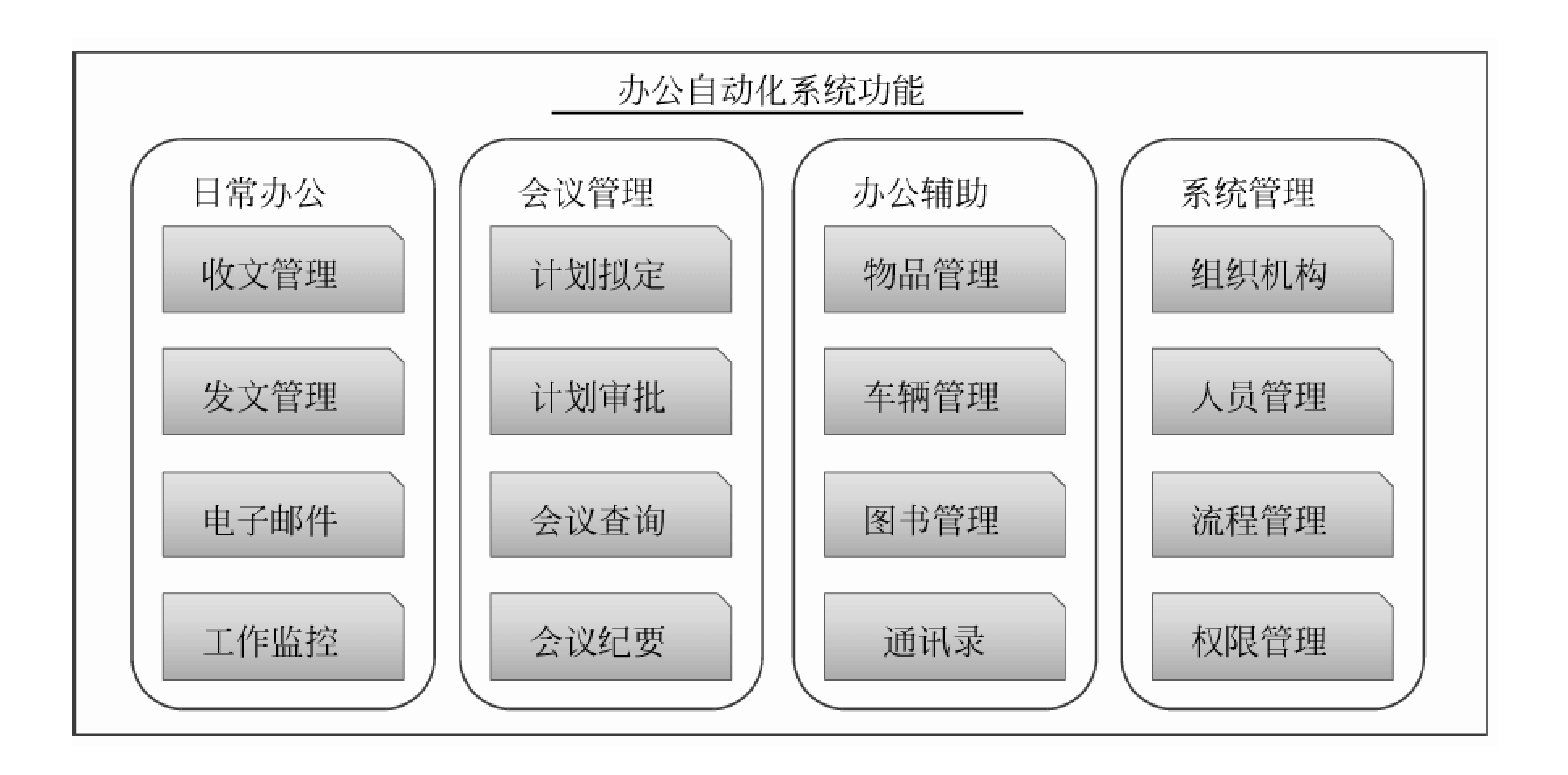


图 1-5-1　办公自动化系统功能

从图 1-5-1 可以看出，办公自动化系统在功能设计时，会划分为若干功能模块，各个功能模块各司其职，完成特定的功能，同时功能之间又相互配合、相互支持。

以产品管理域中产品生命周期管理应用为例，其能力要求与系统功能的对比如表 1-5-1 所示。

表 1-5-1　应用能力与系统功能对比表

N	应用能力要求	对应系统功能
1	提取产品需求	产品设计管理
2	产品建模	产品设计管理
3	提供详细的产品规格	产品设计管理
4	新产品引入	产品开发
5	管理现有产品	产品目录管理
		业务目录管理
		产品绩效管理
6	产品废弃/退出	产品变更与撤销
7	市场与定价战略实施	竞争对手产品管理

可见，应用能力要求与系统功能之间并不是一对一的关系，应用能力是业务人员与技术人员在能力层面达成的共识，而系统功能则是以应用能力为输入的，在模块化、高内聚、低耦合等设计原则的指导下形成的。

能力、应用、功能三个概念从不同侧面定义系统要求。能力从用户需求角度定义系统

要求，业务用户无须关心信息系统的技术实现细节，只需对信息系统提出具体的支撑能力要求。应用则是一系列功能的集合，是从业务需求角度对能力需求的整理，应用使得业务需求与信息系统更加贴近，但又不与技术实现方式捆绑在一起，因此具有一定的稳定性。功能则由信息系统承载，更加具体，有明确的输入输出要求。

1.6　数据：信息社会的永恒记忆

“数据”是经过电子设备采集并存储后的载体，从业务需求到技术实现，通过概念模型和逻辑模型来定义数据及其关系，通过物理模型来实现对数据的承载。

1.6.1　数据定义及其价值

前文探讨了“信息”，把其归为业务层面，即信息是在业务过程中形成的。比如业务人员叙述其工作内容会说“我们这款产品主要针对年龄 18 到 27 岁之间的客户”，那么这就是一条传递产品特征的信息。业务人员会提供很多类似这样的信息，但是这样的“信息”不能原封不动地放到信息系统中，因为信息是自然描述的、发散的，为了使得业务需求中的“信息”能够被信息系统接受，需要在信息系统分析与设计阶段，将这些“信息”转变为“数据”。当然，这里的“信息”并不是前面信息框架中提到的“信息”，而是人类之间为了沟通交流而传递的内容。

从“数据”的字面看，数据包括“数字”和“依据”两层含义，从上面的“信息”例子中，可以抽取出 18、27 这样的数字，同样，笔者认为这就是产品特征定义的“依据”，这个“依据”来源于业务需求。

在信息技术普及的初级阶段，以上说法还勉强说得过去，因为当时信息系统的作用主要是将纸质媒体记录的内容转变为计算机能够记忆的电子信息，通常是将以上信息按照二维表的形式进行存放，计算机的作用更多地体现为对传统媒体的电子化，通过电子化实现信息的共享，提高工作效率。

随着信息技术的不断发展，出现了图片、语言、视频等多种媒体形式，这些媒体同样是信息和数据的记录，当然底层都是 1 和 0 这样的二进制形式，从而使得数据的覆盖范围

更广，当前，我们可以将一切通过电子形式记录的信息统统称为“数据”。

如果将“信息”定义为现实世界中存在的载体，那么“数据”则是经过电子设备采集并存储后的载体。随着信息技术和网络技术的发展，无论是人和社会的活动还是自然环境的变化，都可以以“数据”的形式，以多种媒体形式、不同格式记录下来。如果人们能够利用这些数据，挖掘其中的价值，将会是一件非常有意义的事情。由于这些数据具有规模大、形成速度快、类型多样以及价值性低的特点，业界将其称之为“大数据”。

1.6.2 数据建模与存储

在业务需求分析过程中，概念模型描述了业务对象之间的关系，但概念模型毕竟只是业务侧的一种表述方式，为了支持系统功能的实现，还需要对其进一步的设计。

当前，为了对数据进行有效管理，形成了多种类型的数据库，比如以关系代数为理论基础的关系型数据库、以面向对象为理论基础的面向对象数据库以及面向文档管理的文档数据库（例如 IBM 的 Lotus 数据库）等。由于每一种数据库都有其适用范围，同时每一类数据库都有不同的数据库实现产品，为了使得数据模型能够适应这些情况，保持数据模型的稳定性以及对不同数据库产品的适应性，通常将数据模型分为逻辑模型和物理模型两种类型。

逻辑模型侧重从业务角度来考虑实体/对象之间的关系，不同于概念模型，逻辑模型更加具体和细化，对于关系型数据库，通常采用范式设计方法，根据业务需求的不同采用不同的范式。为了更好地理解逻辑模型，下面举例说明。电信运营商的三户逻辑模型如图 1-6-1 所示。

从图 1-6-1 可以看出，一个客户可以购买企业的多个产品，每购买一个产品就意味着形成一个订购实例，因此客户和订购实例之间是一对多的关系，同样，由于客户既包含购买者，也包括使用者，这就意味着一个订购实例对应多个客户，因此，客户与订购实例之间为多对多关系。通过引入第三范式消除数据冗余，在客户和订购实例逻辑模型之间增加客户订购实例关系逻辑模型，从而将逻辑模型之间的关系变为一对多的关系。

顾名思义，物理模型就是那个最终要填充数据的模型。由于某个具体的数据库会有不同于其他数据库的特性，物理模型需要与具体的某个数据库产品对应，比如数据项类型、长度等在不同的数据库产品之间不一定完全相同，比如整型在 SQL Server 数据库中以 int

表示，而 Oracle 数据库中则以 number 表示。

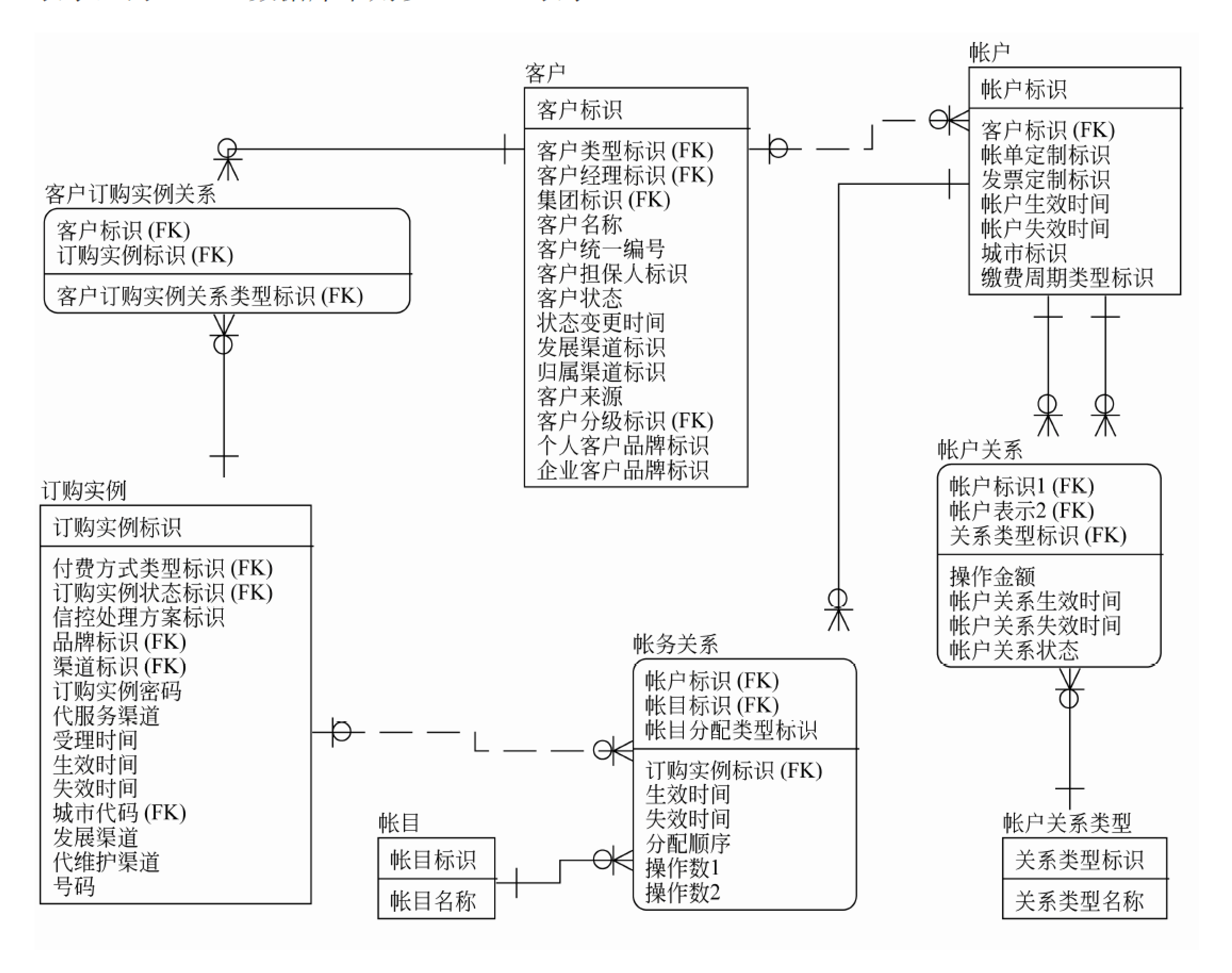


图 1-6-1　电信运营商三户逻辑模型

除了不同的数据库产品在数据类型的定义方面存在差异之外，其他方面也可能存在很多区别，比如数据定义和数据操作语言 SQL，通常是在标准 SQL 的基础上做了扩展，因此在考虑采用用哪一种数据库时，一定要综合数据库产品的价格、特性、售后服务等方面，选择适合的数据库产品。

与业务过程经历了从业务过程→能力蓝图→技术实现的逐步落地过渡过程类似，数据模型同样也经历了从概念模型→逻辑模型→物理模型逐步落地的过渡过程，如图 1-6-2 所示。

从图 1-6-2 可以看出，业务过程与数据模型都是从业务需求逐步过渡到具体实现的，

但都不是一次到位的，都经历了中间一个过渡阶段。之所以这样做的原因是由于专业化分工不同，各参与方的背景知识不同，业务和技术人员对于同一个事物的理解也不一样的，为了消除这种差异以及更好地做好各自擅长的工作，需要通过一个中间媒介来促进共识。对于业务过程来说，需要通过能力蓝图来达成业务人员与技术人员的共识，对于数据模型来说，需要通过逻辑模型达成业务人员与技术人员的共识。

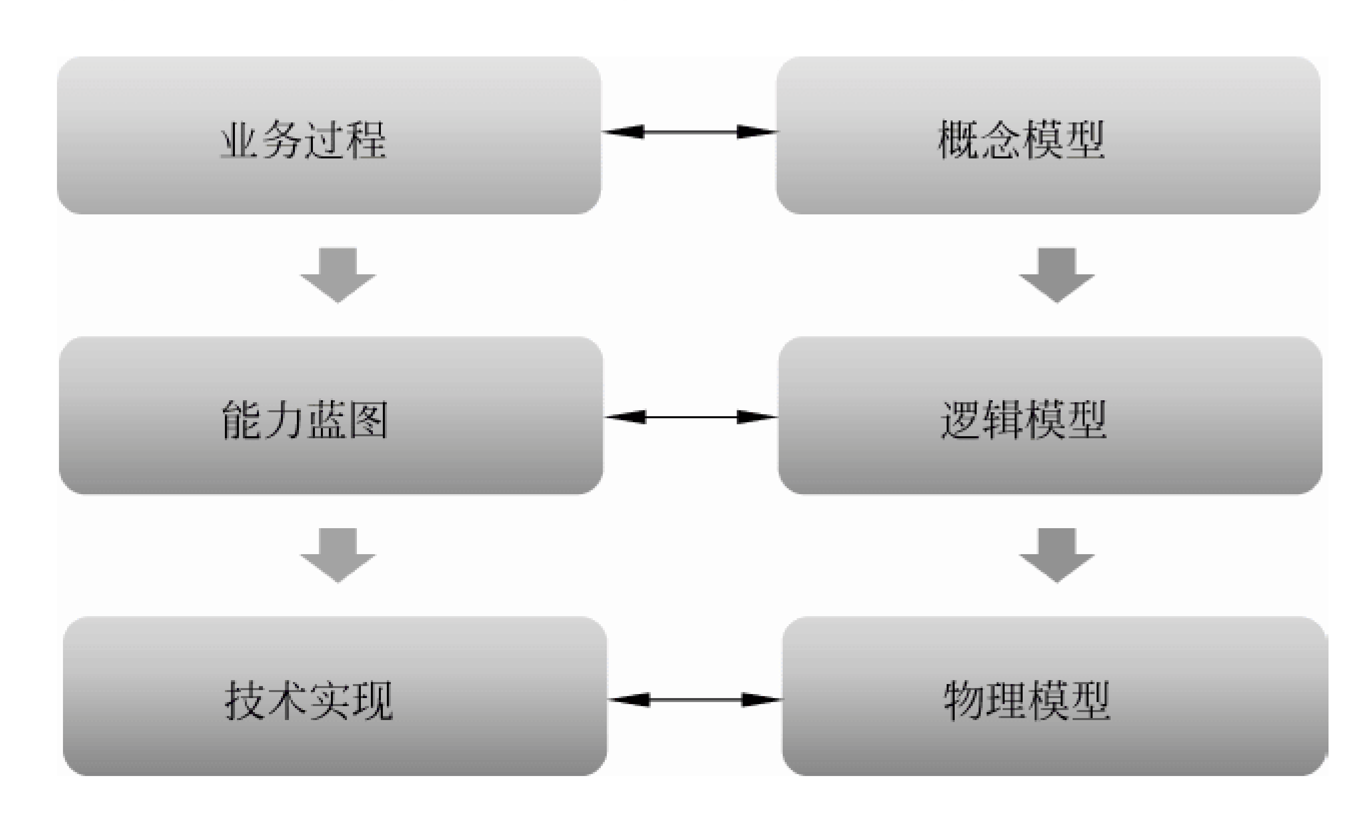


图 1-6-2　业务过程侧与数据模型侧从业务到实现的过渡过程对比

1.6.3　数据的分类

数据可以按多种方式分类。按照数据的媒体类型可以分为文本数据、语音数据、视频数据、图片数据等；按照数据的用途分为生产型数据、分析型数据，生产型数据主要是在生产生活过程中产生的，可能由采集设备采集产生，也可能由人通过使用信息系统时产生，比如某个人在网上购物，那么在该网站上就留下了其浏览、搜索、购物车、下单、支付、投诉等数据，如果某企业在其他企业采购原材料，那么在该企业的采购系统中就形成了企业物资采购的数据，包括采购合同、付款、发票等数据；如果政府部门进行人口普查，那么政府的普查系统中就保留了公民的姓名、年龄、籍贯、出生地、身份证号、家庭成员、教育等数据。

与生产型数据不同，分析型数据以生产型数据为基础，目标是指导生产生活中的各种决策，比如企业通过统计分析，可以找出数据之间的联系、找出规律，从而指导决策。分析型数据的数据基础是对生产型数据进行加工、清洗、转换、丰富等形成的，根据不同主

题的需要，对数据进行建模，以便更好地找出数据背后隐藏的规律，为决策提供参考。

1.7 集成：价值网络时代的整合者

集成的目的就是将整体中的各个部分粘合起来，借助业务服务，可以实现对业务过程、信息、应用、数据、技术等元素的有效集成。

从企业战略到生产再到运营是一个非常复杂的系统工程，为了解决这个复杂问题，通常采用分而治之（divide and conquer）的方法，通过对业务过程、数据模型的分层分类，使得企业能够灵活、快速地响应外部市场的变化。

当问题域被分解后，无论是对业务需求还是信息系统，都带来了另外新的问题，那就是模块之间的集成问题，即如何把这些独立的模块有效地集成起来，以满足特定功能的需要。集成框架的目标就是解决这些问题。

1.7.1 业务层面的集成

为了解决集成问题，可以采用业务服务（Business Services）的方式，将业务过程、信息、应用、数据、技术几个框架中的元素集成起来。

集成框架中的基本元素就是服务，每个服务可能是一个原子服务，也可能是一个组合服务，每个服务可能有自身依赖的服务，也可能作为其他服务的输入，多个服务以价值链思维为导向，可以直观、准确地描述一个业务，可以很好地适应外部需求变化。

按照服务的性质，可以将服务分为三种类型：以任务（task）为中心的服务、以实体（entity）为中心的服务和以效用（utility）为中心的服务。以任务为中心的服务是从动态角度定义服务的，比如“融合业务订单处理”就是一个以任务为中心的服务；以实体为中心的服务是从静态角度定义服务的，比如客户资料查询、账单、详单查询等，都是以客户、账户等业务对象为中心的，因此可以称为以实体为中心的服务。此外，有些服务并不是业务角度能够确定的，这些服务是为了更好地支撑业务的实现，属于公共服务，比如日志服务、异常处理服务等，这些服务就是以效用为中心的服务，相当于任务服务和实体服务的公共设施（utility）服务。

服务描述包括服务的内容、性质、生命周期等，如果具有技术限制或要求，也需要描述服务实现的技术手段。服务表达公式为：服务描述=服务名称+服务类别+依赖服务+支撑服务+服务周期+技术依赖。举例为：融合业务订单处理服务=融合业务订单处理+以任务为中心的服务+融合业务订单接收服务+融合业务订单回笼服务+1 年+JDK5.0。业务服务从定义到实现的过程如图 1-7-1 所示。

从图 1-7-1 可以看出，业务服务是集成的核心元素。业务服务分别来自于以任务为中心的业务过程框架、以实体为中心的信息框架以及以效用为中心的应用框架。当业务服务设计完成后，就可以进行组件的定义和接口的定义了，组件主要用于信息系统内部，组件需要遵循 SCA（服务组件架构）和 SDO（服务数据对象）规范，接口用于信息系统之间的集成。当完成组件和接口的定义以后，就可以进行代码的创建、软件开发、测试等后续实现过程了，最终形成满足业务需求的信息系统。

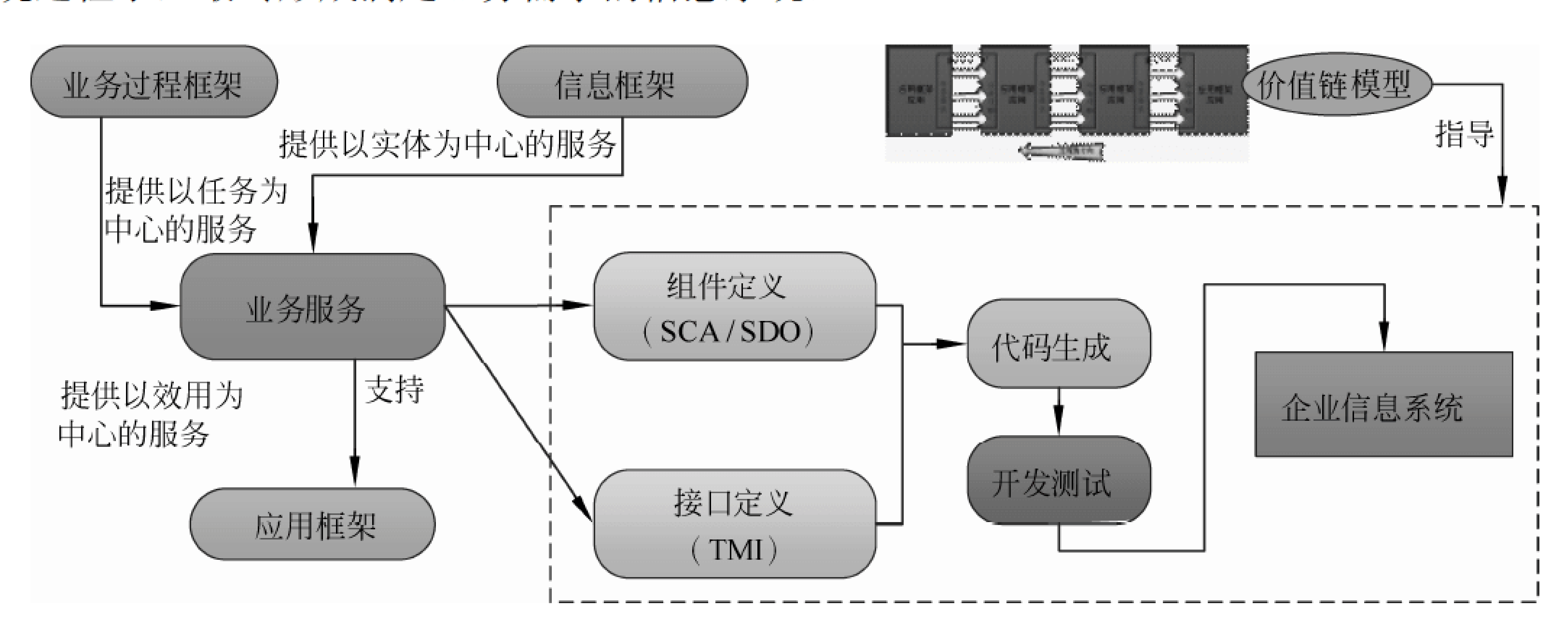


图 1-7-1　业务服务从定义到实现的过程

集成框架以业务服务为中心，一方面是为了构建面向服务的企业，另一方面，采用价值链思维，使得企业可以柔性地适应外部环境，与其他企业进行平滑地对接，快速灵活地适应外部市场变化。

1.7.2　技术层面的集成

技术总是不断发展变化的，1.7.1 中为了解决因为技术变化提出了采用业务服务集成的办法，尽管如此，集成问题最终还要体现在信息系统之间的集成，也就是技术层面的集成。

当然，如果企业只有一个系统就并不存在系统之间的集成问题，但是企业为了提升扩展性和灵活性，通常规划并建设了许多信息系统，这些独立的信息系统分别承载不同的功能。某个业务往往需要通过多个信息系统之间的集成才能实现。比如，电信运营商为了完成某个业务的开通，往往需要集成多个信息系统，比如电子渠道系统、客户关系管理系统、综合计费账务系统、服务开通系统、资源管理系统。将应用分解到多个独立系统最大的好处就是不会因为某个小的功能改变而重新构建所有信息系统，这样可以提升系统的灵活性。

信息系统之间集成有很多具体的实现方式，包括 FTP、Socket、Web Service、DBLink 等，既可以采用企业内部私有协议，又可以采用公共协议。FTP 协议的优点是简单、高效，通常用于信息系统之间批量传送数据文件，缺点是数据质量和数据安全难以保证。如果信息系统之间为单个接口、少量数据的调用，则通常采用 Socket 或者 Web Service 等接口方式，由于 Web Service 为标准化的接口调用方式，实现方式具有与底层实现平台的无关性，因此更容易实现异构（不同操作系统）之间的集成，但是这种跨平台性也是采用在原有协议基础上增加 Header 为代价的，相对于 Socket 接口方式增加了更多的打包与解析动作，进而花费了更多的资源与时间，因此接口效率上会差一些，对于那些实时性较高的应用可以考虑采用 Socket 等私有协议。

当然，网络质量、设备可靠性、应用程序健壮性等因素可能引起数据丢失、安全性、可靠性降低等问题，进而影响到业务提供的质量，因此需要采用一些手段来消除以上不足，比如定期的数据稽核、数据审计等，对传输失败的数据进行重传，最大限度地消除因信息系统集成带来的问题。

随着技术架构的不断发展变化，最近业界提出了“平台+应用”的架构模式，这种模式将多个信息系统公用的支撑功能转移到平台上实现，这样可以基于平台提供的基础功能快速构建新型应用。“平台+应用”模式的典型代表为苹果应用商店（App Store），全社会的开发者只要遵循评估公司制定的开发规范，在苹果公司的软件开发工具包的支持下，就可以自行开发创新型应用，并将其发布到苹果商店之中，这种软件架构模式大大激发了全社会开发人员的积极性和创造性，是对创新能力的进一步释放，是软件开发模式的又一次革命。同样，“平台+应用”的架构模式也降低了系统集成风险，而是将系统之间集成的工作转移到平台层面，通过平台来保障信息交换的可靠性和安全性，大大提供了应用推广的速度和质量。

1.8 技术：改变世界的源动力

> 构建技术架构的目标是保障系统的可靠性、可用性、可伸缩性、高性能以及安全性，分层、组件化和开放是技术架构设计的主要方法。

构建技术架构的目标是保障系统的可靠性、可用性、可伸缩性、高性能以及安全性，此外技术架构还要保障从业务需求到技术实现更好地衔接起来。

在软件出现的早期，通常采用面向过程的分析与设计方法，将信息系统分解为多个功能模块，系统结构为客户端-数据库服务器的两层架构模式，计算逻辑通常在客户端实现，服务器端为专门负责数据存储管理的数据库，典型的语言工具包括 PowerBuilder、Delphi、Visual Basic 等。

随着 Web 技术的发展，技术上逐渐采用三层架构的方式，即浏览器-应用服务器-数据库服务器方式（也称为 B/S 结构，B/S=Browser/Server），应用服务器又可以分为两层：Web 展示层和业务逻辑处理层，为了支持分布式计算等集群功能，Web 层通常需要 Web 应用服务器的支持，业务逻辑处理层则需要 EJB、COM+、CORBA 等分布式组件技术的支持。这种架构模式的优点是将计算逻辑挪到了服务器端，减轻了客户端的计算负荷，客户端则专注于界面展现工作。

但是 B/S 结构模式也有着天然的不足，就是浏览器客户端对于鼠标键盘支撑力度不够，对于那些需要快速记录客户信息的应用（例如客服中心受理系统）显然是不适合的，为了解决这一问题，采用 C/S 和 B/S 混合架构的方式，即客户端内部集成 Web 浏览器控件，服务器端不变，C/S 和 B/S 混合方式综合利用两种架构的优点。当前，如 360、UC 等 Web 浏览器均采用这种架构方式。

1.8.1 云技术架构模式

自从互联网企业的领导者谷歌出现以后，分布式计算、网格计算、并行计算等技术又得到新的发展。为了降低硬件成本，谷歌公司采用 GFS、BigTable 等软件技术，实现了对大量低端机器设备的利用。大量低端配置的主机进行动态资源分配，提高了设备利用率，

同时借助容错机制，保证了数据的可靠性，这就是今天人们经常听到的云计算架构。

云计算时代的到来，改变了传统的 C/S 和 B/S 技术架构。云计算架构将系统分为 SaaS、PaaS 和 IaaS 三层，其典型技术架构如图 1-8-1 所示。

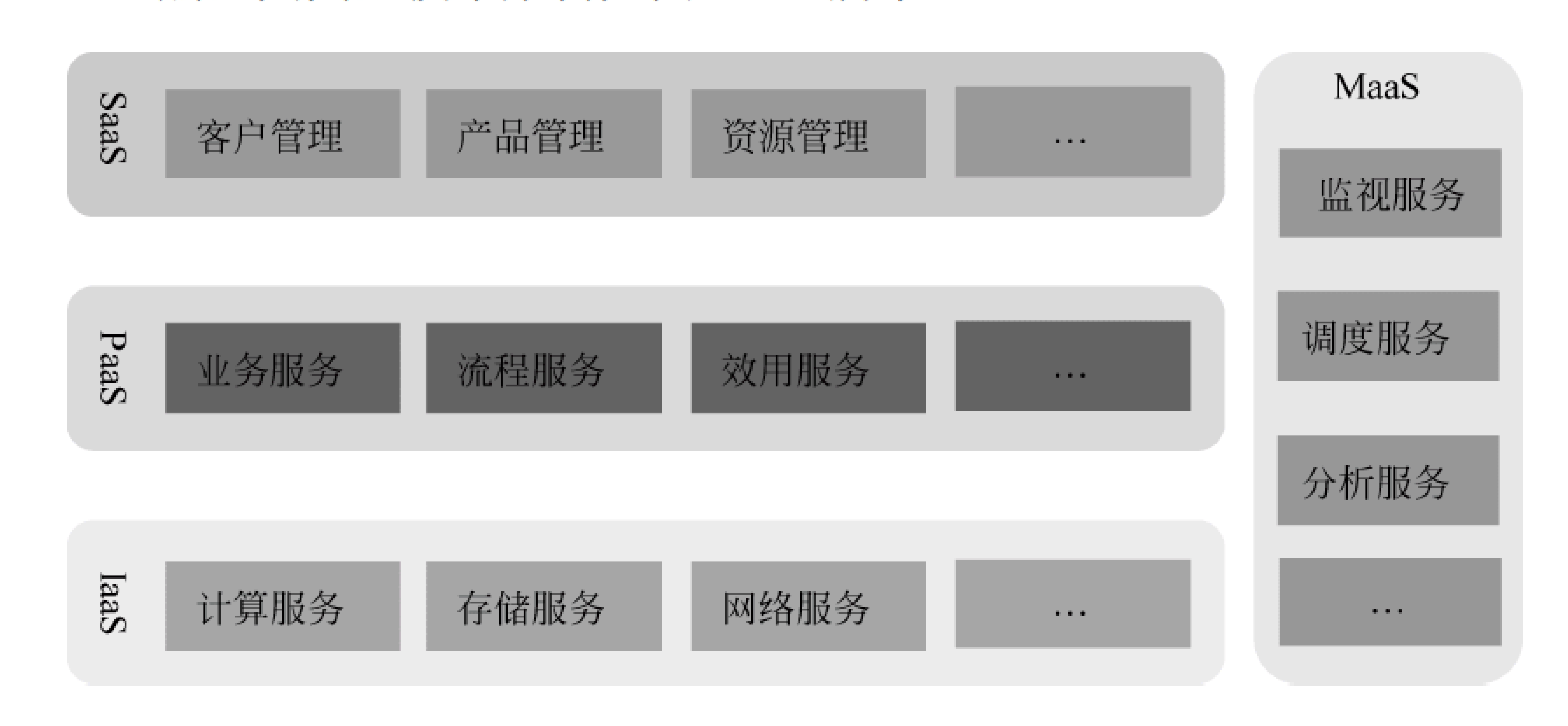


图 1-8-1　云技术架构模式

从图 1-8-1 可以看出，云计算架构主要包括四个部分：软件即服务（SaaS）、平台即服务（PaaS）、基础设施即服务（IaaS）、管理即服务（MaaS）。

软件即服务说明云服务提供的内容为软件应用，对于云应用的用户采用租的方式获取软件服务，好比人们要满足住的需求，不一定非要买，完全可以通过租房的方式实现住房的需求。这种模式对于用户来说非常灵活，如果不想继续使用软件服务则无须继续付费，对于软件服务的提供方，也可以聚集大量有需求的用户，提高软件服务带来的收益。

平台即服务是一种为用户提供“半成品”的服务模式，这种模式可以给客户构建软件服务留有一定的定制空间，在平台服务的支持下，用户可以快速地实现软件服务。好比做饭，买来饺子皮，只要再完成菜馅儿准备、包饺子、煮饺子几个环节，就可以吃到饺子了。因特网服务提供商（ISP）提供的主机服务也是一个典型的平台即服务的例子，只需要将自己开发好的网站应用部署到 ISP 的主机上即可，而无须关心主机设备摆放、公网 IP 申请等事情。

基础设施即服务是最底层的云服务，包括计算服务、存储服务、网络服务等，用户可以根据自身需求选择基础设施服务的配置，就像选择个人电脑一样，不同的是用户无须获取基础设施的所有权，同样采用租赁的方式就可以使用基础设施，无须占用设备空间，不用考虑电力供应等问题。

管理即服务是对以上三种服务进行管理而存在的，通常以上三种层次的服务都是云服务提供商来负责管理的，云服务使用方只关心使用即可。但是，在有些情况下云服务的用户需要自己掌握云服务的运行情况并进行维护，这时云服务提供商可以将对于云服务的管理权限开放给用户，让用户可以自行管理。这好比电信运营商的客户网管系统，企业客户租用电信运营商专线资源并能够查看租用的专线资源拓扑结构及其运行状况，可以根据电信运营商提供的专线服务效果付费，这样透明的服务提供方式是企业发展用户的一种手段。

1.8.2 大数据技术架构模式

生产型信息系统的目标是支撑企业的战略、建设、运营以及企业的管理，其输入为来自企业不同部门的需求，用户通过使用信息系统提供的应用来满足其需求。

生产型信息系统的技术架构通常包括三层：接入/界面/应用层、业务逻辑/平台层、集成/数据层，当用户使用信息系统的应用时，系统接收用户的输入，然后通过业务逻辑层的计算，产生或者读取信息系统的数据。

对于分析型信息系统，与分析型信息系统的数据路径整合相反，如果说生产型信息系统“生产”数据，那么分析型信息系统则“消费”数据。分析型信息系统技术架构包括三层：集成数据层、数据挖掘/平台层、接入/展现层。集成/数据层首先采集和存储来自生产型信息系统的数据，然后经过加工、整合后存储到分析模型之中，最后通过图形、表格等展现方式展现数据分析结果或者将数据分析结果以数据/功能的形式再次注入生产型信息系统之中。生产型信息系统与分析型信息系统技术架构的对比如图 1-8-2 所示。

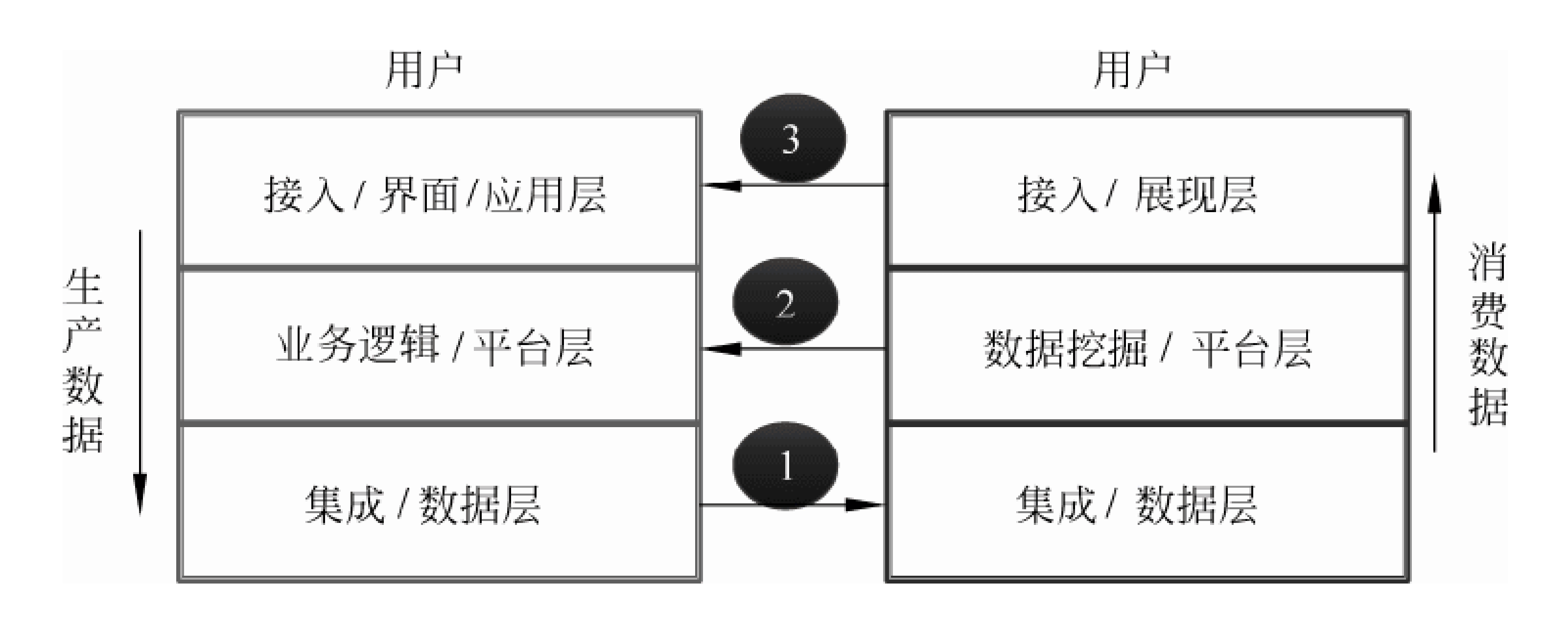


图 1-8-2　生产型与分析型信息系统技术架构对比

从图 1-8-2 可以看出，生产型信息系统与分析型信息的技术架构是类似的，最大的区

别是生产型信息系统生产数据，而分析型信息系统则消费生产型信息系统生产的数据，然后通过整合与挖掘，借助分析模型和算法，发现事物之间的联系和规律。生产型信息系统则是利用分析结果来提高企业生产和经营能力的，人们也可以利用分析后发现的规律，增强认识世界和改造世界的能力。

为了更清晰地看到大数据从采集到利用的过程，下面对大数据技术架构进行设计，大数据技术架构如图 1-8-3 所示。

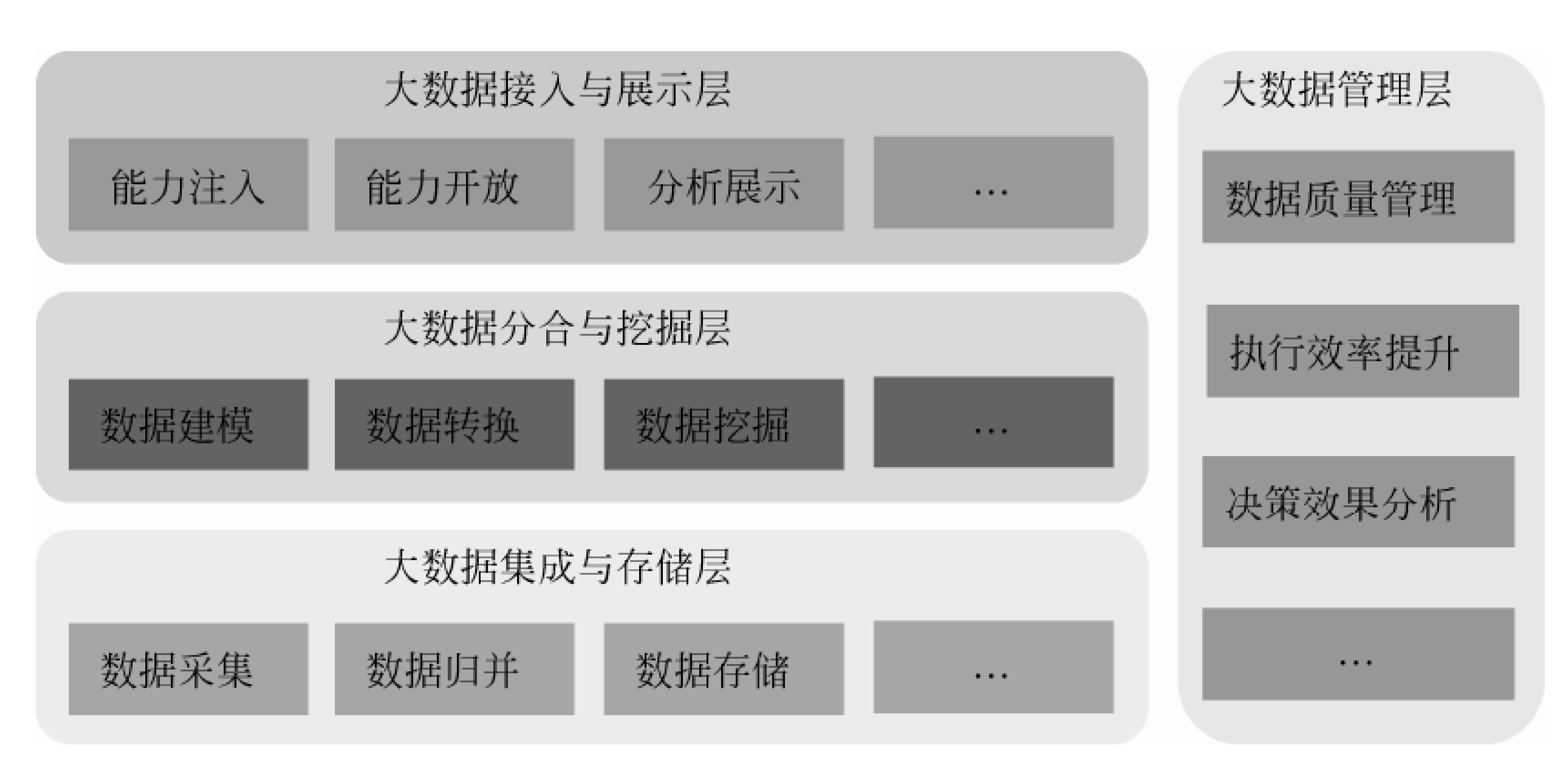


图 1-8-3　大数据技术架构模式

从图 1-8-3 可以看出，大数据技术架构包括四个层面：大数据集成与存储层、大数据分合与挖掘层、大数据接入与展示层、大数据管理层。

1. 大数据集成与存储层

大数据集成与存储层的职能是从各个数据源采集数据并存储到数据库中，可以根据数据的规模采用不同的采集方式和数据库。对于数据规模大并且增量空间不确定的数据需要采用与 Hadoop 类似的分布式数据库，这样可以保证随着数据量的不断增大，数据可以横向扩充。对于数据价值高但是扩展规模可以预期的建议采用传统关系型数据库，这样可以保证能够从多个维度对数据进行挖掘和统计。

2. 大数据分合与挖掘层

大数据分合与挖掘层的职能是找出数据之间的联系和规律。世界是普遍联系的，从自然界或者人类活动中采集的数据反映了自然界和人类的活动，可以借助大数据相关技术和

工具来发现和利用规律。对于企业来说，可以利用发现的规律来提高生产经营能力。之所以称为“分合”，是因为根据应用需要，在数据模型的承载下，通过对于大数据的“分解”与“合并”，形成满足不同应用需要的数据，是大数据挖掘的基本方法。在数据挖掘中通常将“分解”称为“切”，包括切片、切块等。这种思想与财务会计中对于成本费用的“分摊”和“归集”类似，都是通过对操作对象的“微分”和“积分”，满足不同维度的分析需求的。

3. 大数据接入与展示层

大数据接入与展示层的职责是将大数据分析的结果用起来。大数据分析的目的还是“用”，“用”的方法有好几种，一种是直接将分析的结果自动化地植入生产经营的过程中，这是最理想也是最有效率的方式，当然这种方式的缺点是分析的结果可能存在问题，从而做出错误的或者不准确的生产经营决策；另一种方式是为生产经营决策提供“智力”支持，人们可以参考大数据分析结果，结合自身的经验和直觉进行决策，这种情况主要适用于企业战略层面的决策，因为这些决策往往需要企业战略管理人员根据很多年的生产经营经验做出判断，而这些经验往往是没有系统化的数据积累的。

4. 大数据管理层

大数据管理层的职责是对数据使用全过程的监控和管理，包括数据质量管理、数据分析效率分析、数据分析应用效果评价等。

对于数据而言，最关键是要保证数据是真实的、完整的，如果没有好的数据质量作为保障，那么只会产生错误的数据分析结果，形成错误的决策，对企业生产经营造成损失，可见数据质量是大数据分析过程中最最重要的一环。此外，对于大数据来说，如果数据的规模很大，那么如何保证数据的执行效率也是非常重要的。

对于数据分析的效率，不同的应用要求不同，不过一般是效率越高越好，数据分析结果出来得越早越有助于人们快速及时地做出决策，避免因为决策延误而错失市场机会。

执行效果分析对于大数据来说也是非常重要的，因为大数据分析的目的还是要保证能够做出科学合理的决策，由于世界的矛盾性，十全十美的决策是不存在的，但是经过各种因素的考虑之后，需要在利弊得失中做出最优决策。因此，要对大数据分析的效果进行及时评价，如果大数据分析不能达到决策支持的目的，应当尽快寻找支持决策的新方法和途径。

1.9 部署：让飞机平稳着陆

部署是设计方案和系统实现的落地，它将处于不同层级的“硬件”和“软件”有机地结合起来，最终实现可供用户使用的系统和服务。

软件以其神奇的适应性为人类社会提供了各种各样的信息服务，可是，这种“软”的物件最终还是需要“硬”的物件做依托才行，否则软件永远是个想象中的东西。

那么，软件如何才能“落地”呢？这种实现方式业界称之为“部署”（deploy），拿一个简单的个人电脑为例，首先要有电脑硬件，包括主板、CPU、内存、硬盘、声卡、显卡、网卡、键盘、USB 口、显示屏等。有了这些硬件，还要有主板 ROM 芯片中负责个人电脑基本输入输出控制的 BIOS 程序。然后可以开始安装如 Windows 7 这样的操作系统软件，接着就是安装 Office 办公软件、Eclipse 等开发工具软件等。在这里，Windows 7 属于系统软件，Office 属于应用软件，系统软件部署在个人电脑硬件上，而应用软件则部署在操作系统这样的系统软件上。

当然，上面的例子比较简单，只是想说明什么叫部署，在实际的复杂应用中，往往是由网络连接的客户端和服务端组成的，人们通常所说的部署，更多地是指硬件的部署和位于服务端的软件部署，而通常将客户端软件部署称为“安装”。

按照部署的先后顺序，首先是硬件部署然后才是软件部署。硬件部署通常包括网络设备、主机设备、存储设备等的部署，当硬件部署完成后，就可以在硬件上安装系统软件和应用软件了。应用软件、系统软件以及系统硬件的部署层次结构如图 1-9-1 所示。

3	应用软件层	客户关系管理、计费账务管理、客服等
2	系统软件层	操作系统、中间件、数据库、平台等
1	系统硬件层	机架、电源、主机、存储、网络等

图 1-9-1　系统部署层次结构

从图 1-9-1 可以看出，对于信息系统来说，通常是由以上三种层次组成的，需要按照 1、

2、3 的顺序自下而上部署，这样才能形成最终用户能够使用的应用。

1.9.1 部署的不懈追求：5 个不变

当前，随着云计算和大数据时代的到来，对于部署的架构模式提出了新的要求，但是云时代的部署模式本质上与传统的部署模式是一样的，它们都是为了满足系统在可靠性（Reliability）、可用性（Availability）、可伸缩性（Scalability）、高性能（Performance）以及安全性（Security）这 5 个方面的要求。

1．可靠性

可靠性是衡量信息系统服务质量的关键指标，一个不可靠的信息系统会导致服务中断，客户体验水平会大打折扣，会直接破坏企业的服务形象，为企业带来或大或小的经济损失。试想，如果某电子商务网站经常因为信息系统的不可靠而无法登录购物系统，谁还去这家电子商务公司购物？如果某家企业的 ERP 系统经常无法使用，那么企业内部员工如何办公？如果公检法机关的电子政务系统无法使用，那么国家公务人员如何执法？可见，信息系统的可靠性是极其重要的。

从信息系统的部署角度看，首要目标是要保证信息系统的可靠性。为了实现这一目标，需要从部署对象的各个层次来保障可靠性，通过灵活的部署方案实现信息系统的整体可靠性。

1）系统硬件层

系统硬件层包括电源、主机、存储、网络等，属于信息系统的基础设施部分。要保证系统硬件层的部署对象的可靠性，通常采用增加资源的方式，通过硬件设备的互相备份达到服务不中断的目的。

电源设备是信息系统可靠性最核心的设备，如果没有可靠电源的供给，所有 IT 设备将是一堆废铜烂铁，因此保证可靠的电源供给是极其重要的。保证电源可靠供给的方式主要包括两种：第一种是采用双路供电，这样当一路供电出现故障时，另一路供电可以接上。另一种方式是采用不间断电源（UPS），保证在全部电源输入出现故障后，仍然有蓄电设备持续供电，并且可以实现电力供给的无缝切换，保证信息系统不会宕机。当然，UPS 供电方式仅仅是一种临时的电源供应方式，因为通常 UPS 的蓄电容量有限，电力供给的时间也有限。一般来说，电力供给出现故障后在一定时间段内就能够恢复，这样当信息系统有正

常的电力作为供给后，UPS 设备又变为一种备用方式，充满电后继续备用。因此，通过在线供电和离线供电的配合，可以实现电力供给的连续性，保证了信息系统能够不间断地提供服务。

对于主机来说，保证其可靠性的方式通常是建立服务器集群，通过集群来消除因为某个或者某些主机宕机引起的单点故障。在一个服务器集群中，多个主机运行同一计算逻辑，这样可以保证计算功能能够顺利地切换到正常运行的主机上。在主机形成的集群中，可以设置主机之间为互相备份的方式，既可以采用一对一的备份方式，也可以采用一对多的备份方式。

对于存储来说，通常是采用冗余磁盘来保证数据的可靠性的。所谓磁盘是通过磁介质来记录数据 0 或者 1 的，这些磁介质因为受碰撞、外部磁力、温度、湿度等外界环境影响而可能出现消磁现象，进而影响数据存储的可靠性。同时，无论对于企业或个人，数据又是最最核心的东西，因此如何保证数据的可靠性就变得非常重要了。保证存储可靠性的方式主要包括两种：第一，通过构建磁盘冗余阵列来保证数据的可靠性，这就是人们经常听到的 RAID（Redundant Array of Independent Disk），RAID 分为 RAID0、RAID1、RAID5、RAID 0+1 等多种方式，不管哪种方式从原理上都是用冗余磁盘空间来换取可靠性的，也就是用的磁盘空间越大，磁盘内容备份就越方便，数据丢失的概率就越低，数据存储的可靠性就越高。第二，考虑到数据存储设备的成本和适用范围，通常采取多级存储（Cache、内存、磁盘、磁带）的方式来设计存储架构。一般来说，存储设备的性能越高、存储空间越大，存储设备的价格也就越高。从应用的角度看，并不是所有应用都需要性能和价格都高的设备的，因此需要根据应用的实际需要对存储介质的类型、数据存储设备的容量等进行规划设计。

对于网络来说，实现可靠性的方式是双路由，就好像人们平时开车，可以走 A 路线，也可以走 B 路线，当 A 路线因为修路或者交通事故等原因封路后，可以走 B 路线，因为有两种路线或者多种路线选择才不会影响人们出现。保证网络可靠性的方式也是采用多路由，基于 TCP/IP 的互联网中的路由策略就是不断寻找可以通行的网络路径，从而保证数据包能将交付到目的地。

2）系统软件层

构建于系统硬件层之上的系统软件层的目的是为了管理系统硬件资源并便于应用软件的实现。为了提高应用软件的开发、测试、部署效率，形成了诸如操作系统、中间件、数据库等系统软件，当然，也存在如何保障系统软件的可靠性问题。

对于操作系统这样的系统软件，由于经过了严谨的设计以及多年的修补和完善，可靠性大大增强，例如当前业界主流的Linux操作系统，当然也不能排除操作系统仍旧存在Bug，解决操作系统可靠性的办法就是通过实时监控和集群的方式，当发现问题后就进行修复和完善。例如微软的 Windows 操作系统不定期发布的补丁就是在发现并解决 Windows 操作系统存在的问题后形成的。

对于中间件，分为交易中间件、Web 中间件、消息中间件等多种类型，中间件的出现也是为了保证信息系统有更好的可靠性，试想如果没有中间件，每个应用软件都要从头开始，一定会存在许多 Bug。介于操作系统和应用软件之间的中间件专注于某一个特定领域解决特定的问题，形成了可复用的软件功能和组件。由于中间件具有先进的架构设计、高水平的开发以及长期的应用实践，因此具有很高的可靠性。另外，为了保障信息系统整体的可靠性，系统软件可以构建成一个集群，集群内部的某些中间件是其他中间件的运行副本，保证在某些中间件节点出现故障后信息系统仍旧能够正常地提供服务，同时集群方式也可以提高信息系统的总体吞吐量。

数据库也可以认为是提供数据管理服务的中间件，由于数据库中保存了大量有价值的数据，保障数据库的可靠性就显得更为重要。为了保证数据库的可靠性，Oracle 数据库采用了实时应用集群（Real Application Clusters，RAC）技术，RAC 作为集群技术的一种，可以很好地实现数据库的可靠性。

3）应用软件层

相对于系统软件，应用软件更多地体现了业务的特殊性，应用软件大多是个人和组织从应用软件开发商处购买或者从软件服务提供商处租用得到的。

应用软件的类型非常多，企业普遍使用的应用软件包括客户关系管理系统（CRM）、客户服务系统、企业资源计划系统（ERP）、办公自动化系统（OA）、供应链管理系统（SCM）等。

应用软件的可靠性一方面需要以先进成熟的系统软件为基础，另一方面也可以通过单元测试、集成测试、系统测试、用户接受性测试（UAT）等方式和手段，不断优化和完善软件代码实现。应当在软件代码中通过异常处理和日志记录的方式来定位应用软件中出现的错误，保证应用软件的可靠运行。

2. 可伸缩性

信息系统不但要面向现在，还要面向未来，应当具有信息系统容量的可扩充性，在互

联网时代，弹性的、可伸可缩的系统架构成为信息系统部署设计阶段重要的考虑因素。

对于单个系统硬件，由于硬件物理特性的限制，保证可伸缩性的唯一方法是提前预留可以扩充的插槽和端口，比如目前流行的刀片式服务器，就可以根据需要增加或者删减服务器，服务器好比“刀片”一样，可以方便地插入“刀箱”之内。

单个硬件的扩充能力毕竟是有限的，为了解决动态的资源扩充问题，业界通常采用集群架构来实现系统硬件的横向扩展，集群内的资源通过集群管理软件来进行配置和管理。当前，虚拟化、云化等技术就是将资源使用和资源分配相隔离，动态地进行资源调度，使得信息系统具有良好的可伸缩性。

3. 可用性

可用性是从用户对于系统服务的角度提出的。简单地看，可用性包括可用和不可用两种类型，不可用就是不可以使用，当然是用户不能接受的。为了衡量信息系统服务的质量，系统服务提供方和使用方之间可以通过可用性指标作为支付的评判标准。

信息系统可用性的实现是一个系统性工程，因为对于应用的使用者是不关心系统硬件出现问题还是系统软件出现问题的，用户所要求的就是软件总是可以使用的。

从部署的角度看，保证可用性的方式是进行充分的集成测试和系统整体性测试，发现并解决影响系统可用性的因素。

4. 性能

与可用性类似，性能也是用户能够直接衡量的一个方面，当然，不同的应用对于性能的可接受程度是不一样的。对于性能的具体要求，通常是要求性能在一个可接受的时间范围之内，比如 1 天或者 5 秒之内。

要实现系统的高性能，同样需要信息系统各个方面的努力。CPU 的频率、磁盘存取速度、IO 处理能力等都是影响系统性能的重要因素。对于需要应多大规模并发用户的信息系统，同样需要集群部署的方式来提高系统的整体性能。

系统性能与集群规模、数据存储方式、程序算法等都有关系，在大数据时代，海量数据处理与快速展现分析结果之间的矛盾更加突出，使得解决系统的性能问题变得更加具有挑战性。

5. 安全性

安全性也是部署时考虑的重要因素，安全性与可靠性是不同的，安全性主要是防止未经授权的用户对于系统资源的访问。

提高系统的安全性需要从技术手段和规章制度两个方面做起，本文主要从技术角度来分析如何部署才能保障系统具有高安全性。

首先，安全问题产生的根源通常是在信息系统访问客户端，正所谓“病从口入”，因此如何控制客户端引起的安全是需要首先考虑的问题。对于可控可管的客户端，应当在客户端部署代理软件，保证客户端是在预先设定的安全策略控制下使用的。此外，应当从网络层面进行控制，对于接入网络的客户端进行安全认证，将不符合安全要求的用户阻挡在网络边缘。

其次，对于传输和存储的数据进行加密，保证传送和存储的数据不被第三方窃取。

最后，部署 4A（账户、认证、授权、审计）服务器，在信息系统的应用层面进行安全控制，及时发现并处置违规的用户。

1.9.2 部署的好伙伴：配置管理

随着系统硬件、系统软件以及应用软件的不断部署实施，数据中心的规模会越来越大，为了保证系统的高可靠性，数据中心甚至采用跨地域的方式来实现，当系统出现故障后，如果不能实时掌握软硬件部署的结构和运行状况，就难以快速地完成故障恢复。为了解决这一问题，迫切需要对系统软硬件资源的配置关系进行管理。

配置项是配置管理的基本单元，配置项根据管理的需要可大可小，可以是一台机器设备，也可以是机器设备中的一个部件。配置项记录了自身的属性以及与其他配置项的关联关系，比如配置项为机架的属性通常包括机架的位置、可用空间、已用空间、安装时间、责任人等，配置项为服务器的属性，通常包括服务器所在机架、电源消耗、厂家型号、硬件配置、安装时间、资产原值、折旧年限、运行状态、装载的软件、连接的网络端口等。

为了保证配置项数据的准确性，要求在系统部署完成后及时更新配置项属性。配置项信息的录入分为两种：一种是类似于机房、机架等无须运行的设备，需要完全依靠人工录入的方式完成；另一种是类似于服务器、存储设备、网络设备等可以运行的设备，可以在

安装完成后通过手工的方式进行数据的初次录入，然后再借助运行后的设备管理软件进行配置上报，对比两次采集的配置项属性，校正存在问题的配置项，提高配置数据的准确性。

对于大中型的数据中心，配置项之间的关系是非常复杂的，因此要求支持图形化方式的配置管理，直观地展示配置项属性及其运行状况，帮助运行维护人员快速发现并修复存在的故障，帮助数据中心规划人员实时掌握 IT 资源使用情况，以便及时的补充资源，保证系统在正常的负荷范围内运行。

1.9.3 事务型应用系统部署架构

下面是某企业使用笔记本电脑、台式机、手机等接入设备，通过因特网访问企业数据中心的部署架构设计方案，如图 1-9-2 所示。

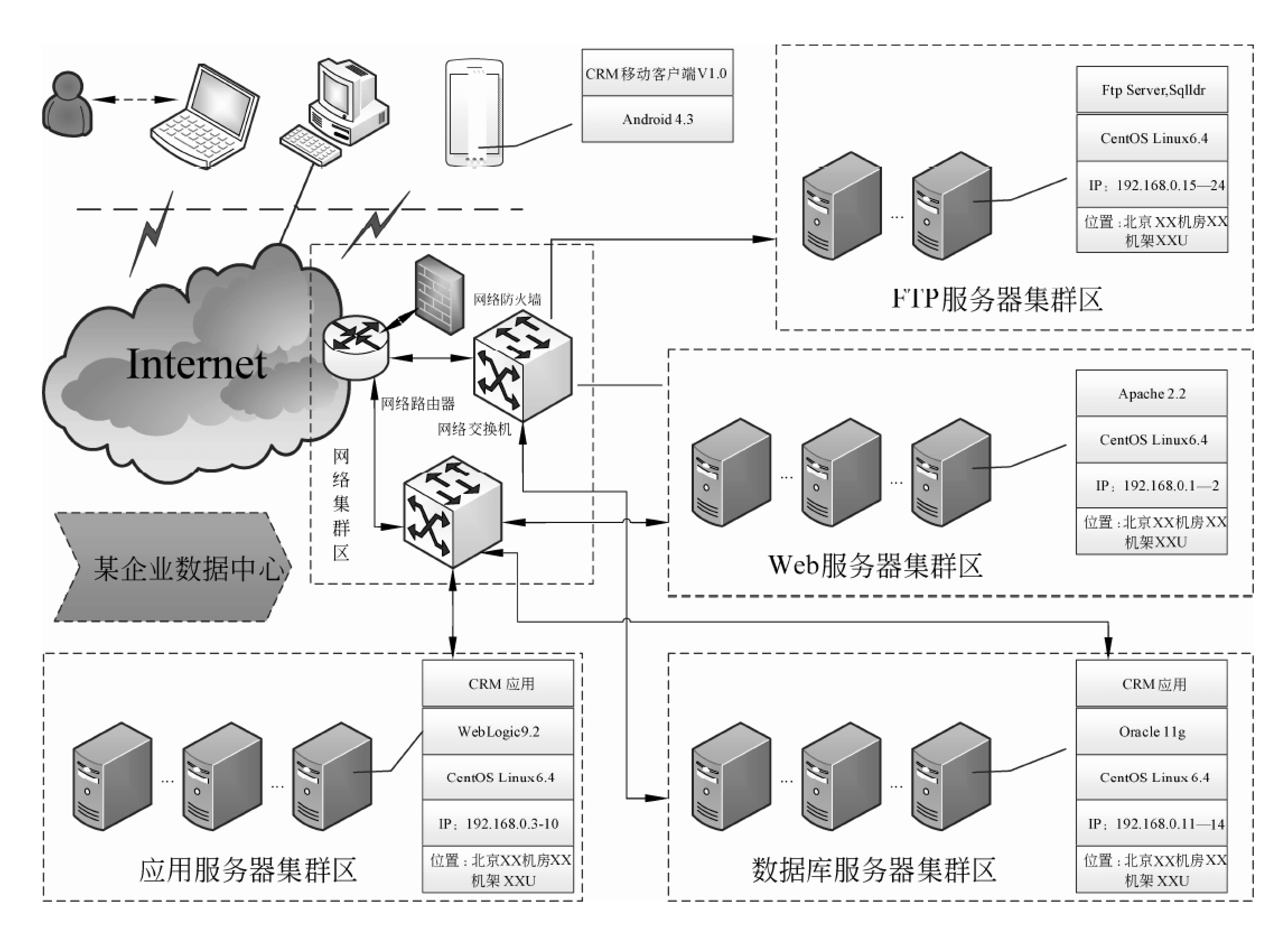


图 1-9-2　某企业 CRM 系统部署架构设计方案

从图 1-9-2 可以看出，企业 CRM 系统分为客户端和数据中心两部分。

CRM 系统客户端分为笔记本电脑、台式机、智能手机三种接入方式，其中智能手机的部署方案为在 Android V4.3 操作系统上部署 CRM 手机客户端软件 V1.0。

CRM 系统数据中心分为 4 个区，即网络集群区、FTP 服务器集群区、Web 服务器集群区、应用服务器集群区、数据库服务器集群区。网络集群区包括网络路由器、网络防火墙和网络交换机，FTP 服务器集群区包括多个 FTP 服务器，起到接收上传文件和导入数据的作用，Web 服务器集群区部署 Web 静态网页并起到 Web 服务负载均衡的作用，应用服务器集群区提供 CRM 应用逻辑的处理，数据库服务器集群区负责存取数据和管理数据。

系统部署结构中体现了从硬件支撑到软件部署的层次依赖关系，最下层为硬件设备及其安装的位置，上层分别为操作系统、中间件、应用软件，这样就可以清晰地看到软件与硬件的依赖关系，有助于项目后期完成系统的维护工作。

由于篇幅限制，图 1-9-2 中没有绘制数据中心的容灾部分，可以认为各种集群区都具有机房内部、同城以及异地容灾的能力。

1.9.4 分析型应用系统部署架构

以上是面向操作的事务型应用在某企业数据中心的部署架构方案。与支持企业事务型应用的部署方案不同，支持企业分析型应用的部署方案有着自身的特点。某电信运营商移动上网记录大数据应用的系统部署方案如图 1-9-3 所示。

从图 1-9-3 可以看出，分析型应用系统架构设计需要考虑数据采集、数据上传、数据装载、数据交换、数据查询等环节。在大数据应用系统部署的不同阶段，主要完成的任务包括：

- 数据采集阶段。首先，移动用户通过手机借助电信运营商的通信网络和互联网访问 OTT 应用，比如腾讯 QQ、新浪微博、淘宝网等。其次，在移动用户访问 OTT 应用的时候，通信网络中的 GGSN、SGSN 等网关设备会记录下移动用户的上网行为，包括上网时间、上网时长、上网流量、应用 URL 等。最后，采集设备将抓取到的数据包存入上网记录文件后，上传到 FTP 服务器。
- 数据装载阶段。首先，FTP 服务器通过 vsftpd 服务器接收上传的文件并将其放入指

定的目录。然后，通过 ETL 软件将数据装载到 Hadoop/HBase 分布式数据库集群中。Hadoop/HBase 分布式数据库集群加载数据的方法为：客户端根据主服务器（Master Server）的 NameNode 找到装载文件需要存放的区域服务器（Region Server），通过区域服务器将数据存放到 DataNode。其存储原理类似于 Linux 的虚拟文件存储，即将所有设备看作一个文件，而 NameNode 则是管理文件的索引。

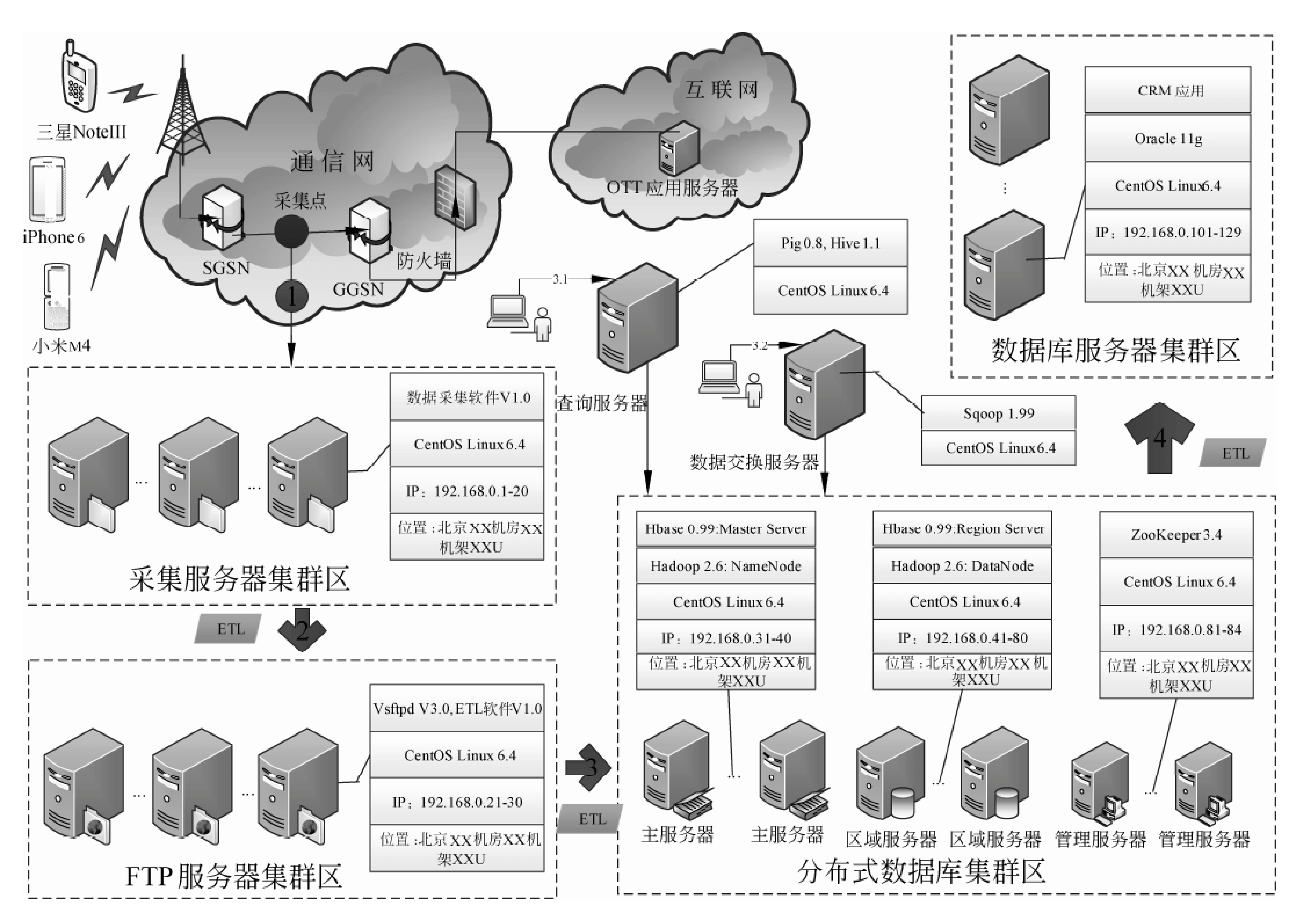


图 1-9-3　某电信运营商大数据中心系统部署总体架构

- 数据查询阶段。分布式数据库的查询方法不同于关系型数据库，可以借助开源工具 Pig、Hive 使得分布式数据库的查询变得更加简单。
- 数据交换阶段。如果存储在分布式数据库集群中的数据需要用于进一步的统计分析，那么需要通过数据交换工具 Sqoop 将其存入关系型数据库 Oracle 集群中，由于关系型数据库对于数据的存取性能受限于数据的规模，因此可以将初步汇总后的数据存入关系型数据库。

1.10 安全：都是开放惹的祸

坚持开放就必然会带来安全问题，可以沿着系统架构的"云+管+端"思路来分析引起安全问题的根源并提供整体安全解决方案。

互联网因为其开放特性使得全球信息可以自由流动，全球资源可以动态调配，从此世界变得"平坦"了。可是，互联网也为这种"开放"付出了代价。

本节首先分析开放带来的问题，然后再基于云管端模式分析安全管理的应对策略与方法。

1.10.1 开放的价值与代价

从历史上看，开放促进了人类社会的发展进步。贸易的开放促进了商品经济的发展，文化的开放增进了不同民族的相互理解，而技术的开放则以信息为载体，以网络为支持，实现了全球资源的重新配置。从这个角度看，由于技术的开放引起的信息流动导致了全球资源的归集与配置，信息犹如货币，哪里需要就流向哪里，从而实现了资源的动态配置。

1.10.2 云管端模式下的安全管理

在云计算时代，将信息系统的承载结构划分为"云+管+端"三个部分，因此对应的安全管理也可以从这三个部分进行分析。基于"云+管+端"的安全管理框架如图 1-10-1 所示。

从图 1-10-1 可以看出，云+管+端实际上是从 IT 服务供给到 IT 服务需求的一种消费模式，云、管、端的每一个部分又可以分为应用软件、系统软件和系统硬件三个支撑层次，两种结构交叉形成的矩阵块就是安全管理的对象，针对不同安全管理对象的安全策略与方法如下：



图 1-10-1　信息系统安全管理框架

1. “云”侧安全策略与方法

“云”侧是 IT 服务的提供方，例如 CRM、ERP、SCM 等提供软件服务的信息系统，分为系统硬件、系统软件以及应用软件的安全管理。

1）系统硬件安全管理

对于像服务器、存储设备等硬件设备，安全性管理与可靠性管理是同样一件事情，对于安全来说，通常是指软件引起的安全问题。可以采用双击互备、容灾等方式来保障硬件级的安全可靠性。

2）系统软件安全管理

系统软件用于实现特定功能的安全管理，例如主机防火墙、病毒查杀、漏洞检测等，系统软件通常由第三方企业提供，企业需要购买通过权威部门安全认证的系统软件，未经安全检验的系统软件可能会成为注入的木马程序。

3）应用软件安全管理

在应用软件层级，主要是解决因为用户行为而导致的安全问题。例如，未经授权的用户使用了系统的功能、窃取了系统内部的数据，企业内部的系统用户访问了不应当访问的数据，从而使企业带来经营风险甚至经济损失。

以上安全风险的防范可以通过事前认证和事后审计相结合的方式解决，对于未经授权

的用户一定要阻止在系统之外，同时对于授权但是违规操作的用户可以通过事后审计的方式找出来。

2.“管”侧安全策略与方法

“管”位于 IT 服务提供方和 IT 服务消费方的中间位置，是联通两者的通道。通信网、互联网等各种网络属于管道，网络内部的通信方式包括光通信、微波通信、短波通信、超短波通信、卫星通信等，按照传播介质分为有线通信和无线通信两种类型。

负责通信的设备包括传输设备、交换机、路由器、集线器、网关设备等，为了保障“管”的安全，通常在网络的出入口处设置网络防火墙、入侵检测设备，以便将来自网络的安全问题阻挡在网络防火墙之外。

3.“端”侧

“端”位于 IT 服务消费方，例如桌面电脑、平板电脑、手机等接入终端。由于终端是引起网络安全的起点，因此对于企业内部的可控终端设备可以采用安全管理系统的方式来设置对外访问权限，对于终端自身的安全，可以采用杀毒软件来实时监控终端系统。企业可以考虑采用桌面云的瘦终端方式简化终端设备，在云端集中管理，对于客户服务这样仅仅具有受理功能的信息系统可以优先考虑采用桌面云技术实现。

1.11 治理：没有规矩不成方圆

治理是对业务、应用与技术的管理，通过组织、人员、流程来保障，由于操作型应用与分析型应用的特点不同，治理重点也不一样。

没有规矩不成方圆，同样的道理，再好的企业架构模式如何没有好的治理方法和手段也会沦为空谈。

到此为止，按照构造房子的方法已经完成了这座神奇的小房子的构建，那么房子是不是符合要求，是不是能够入住以及入住后出现问题怎样解决等还需要通过有效的治理来完成。

从治理的目的来看，无非是将这种企业架构的各个部分良好地衔接起来，当房子的某

个部分出现问题也能够及时修复问题。单纯从 IT 治理的角度看，IT 治理的目标就是保证企业信息系统平稳可靠地运行并能够满足性能、安全、扩展等要求。从 IT 治理的范围看，包括对生产型应用的治理和分析型应用的治理，包括业务应用、平台以及基础设施三个层面的治理，此外，安全管理也是 IT 治理的重要内容。

生产型应用治理的重点是保障信息系统高效、可靠、安全地满足企业生产经营需要，通常是采用良好的治理流程作为保障，比如服务台负责收集信息系统在支持企业生产运营过程中产生的问题，再对问题进行分析判断并转发到后台由不同的人员或者系统来处理。后台又可以按照专业分工，分为应用、平台、基础设施等不同系统维护角色，通过这种前后台流程的协同，及时发现和处理信息系统运行中产生的问题。

分析型应用治理的重点是数据质量管理、元数据管理、数据生命周期管理以及隐私保护。分析型应用就是本书所说的大数据应用。

对于大数据应用，数据就像大楼的地基，数据质量是大数据应用对于企业决策支持的关键，因此大数据治理最重要的是要保证数据的质量，包括数据的准确性和完整性。引起数据质量问题的原因很多，数据处理的各个环节都可能引起数据质量问题，数据处理的环节包括采集、传输、导入、集成、清洗等。

用户在使用生产型应用时，无需关注信息系统内部数据的定义，只需关注使用的功能是否满足要求即可。大数据应用则不同，数据是大数据应用形成的起点，因此首先要掌握数据的定义，才能够谈如何利用数据，对于数据的定义就是元数据。元数据是定义数据的数据，它说明了数据的类型、长度、处理过程、过程方法等，有了元数据，就好像有了一本关于数据定义的字典，无论数据存放在什么地方，只要有元数据，就可以理解和使用数据，开发各种各样的分析型应用。

元数据治理的方法是将数据内容和元数据同时保存，保证元数据和数据内容是对应的，就好比字典的目录中标题的页码与页码对应的内容一致一样。

与企业的产品、客户、资源一样，数据同样具有从形成到消亡的生命周期，在数据生命周期的不同阶段，需要采用不同的数据管理方法。

数据仓库之父比尔 • 恩门在著作 DW2.0 中将数据存储分为 4 个区：交互存储区、集成存储区、近线存储区、归档存储区。交互存储区存储新数据，对于数据操作的响应时间通常在几秒之内；集成存储区的数据来自于交互存储区，存储 1 天或者 1 个月的数据，这部分数据主要满足企业在线分析型应用需求；近线存储区主要存储 3~4 年的数据，主要是满足企业更长期的数据分析应用；归档存储区存储 5 年以上的数据，这部分数据通常是为了

满足政策法规的要求而存储的，通常会很少访问归档存储区的数据。

隐私保护是大数据治理的重点和难点，隐私保护主要是因为数据中具有侵犯个人或者组织隐私的信息，这与数据开放的要求通常相悖，企业可以采用提供匿名数据、统计数据等方式规避数据开放带来的隐私问题。

1.12　本章主要内容回顾

企业架构是企业高效运营的基础，企业架构应当能够很好地支持企业发展战略落地实施，敏捷地适应企业内部和外部环境变化。

企业架构就像一座设计严谨的房子，具有各司其职并且相互连接的构件，可以很好地适应外部环境变化，可以实现从业务到技术的有效衔接。

按照自上而下、动静分离、业务与技术分离的设计方法架构企业，将企业划分为 10 个既相互独立，又相互联系的部分。

在业务层面，按照动静分离的方法，分为业务过程架构和信息架构。业务过程架构属于业务中“动”的部分，而信息架构则属于业务中“静”的部分，业务执行过程中产生信息，信息是业务过程的载体。

在系统层面，按照动静分离的方法，分为功能架构和数据架构。功能架构属于系统中“动”的部分，而数据架构则属于系统中“静”的部分，系统功能执行过程中产生数据，数据是系统功能的载体。

应用架构是业务与技术之间的桥梁和纽带。从业务角度看，应用体现了业务对于系统的能力要求；从技术的角度看，应用体现了系统需要具备的能力。

如果说应用架构是业务与技术之桥，那么集成架构则是业务与技术之间的“粘合剂”。集成架构将业务过程架构、信息架构、功能架构、数据架构集成到一起，使得业务与技术既可以按照动静分离原则分别设计，又可以通过集成架构重新成为一个整体，体现了企业架构的灵活性。

技术架构可以让想法变为现实。复用性和灵活性是技术架构追求的主要目标，以保证系统能够快速地适应外部变化。软件内部分层可以实现软件功能的专业化分工，通常系统可以划分为界面层、业务逻辑层和数据层；组件化是提高复用性的重要手段，尽量将软件单元封装成特定功能集的组件；在移动互联网时代，技术架构特别要遵循开放性原则，通

过引入更多的开发者力量，提升软件的集成能力，促进软件应用的创新。

如果说技术架构侧重于对软件系统的结构设计，那么部署架构则是侧重于软件与硬件的结合。“一个篱笆三个桩，一个好汉三个帮”，在云时代，更加需要集群作战，因此部署架构主要关注集群中各个节点如何配合，以实现系统的可靠性、可伸缩性、可用性、高性能以及安全性。

安全是矛盾的另外一个方面，就好像有正义就有邪恶一样。系统安全问题通常是由于系统开放引起的。与系统架构相对应，系统在“云”、“管”、“端”都会存在安全问题，因此安全架构的设计目标就是要保证“云-管-端”的安全。在“云”端，主要是解决信息安全问题，在“管”端，主要是解决网络安全问题，在终端侧，主要是解决接入安全问题。

治理架构强调系统性地看待问题和解决问题。当新增需求、需求变更、服务中断或者不可用时，企业应当通过科学高效的管理制度和流程予以解决。治理架构要求系统具备良好的自我诊断和修复能力，提高机器的智能化水平，实现人与机器之间的有机配合。企业需要综合职能、过程、全生命周期三种思维方式分析和设计治理架构。

第 2 章
联姻：当企业架构爱上大数据

当在企业架构方法论的指导下，自顶向下、从前到后完成了系统化的架构设计以后，就应当为其注入能量，让其焕发青春了。

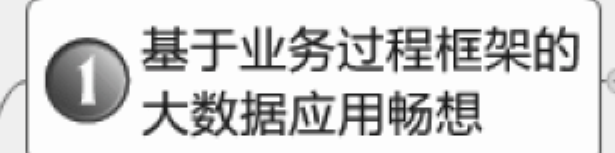
那么，如何才能实现这一目标呢？结论就是借助大数据。虽然以企业战略为指导完成了企业的架构设计，但它还仅仅是一个空架子，长得很丰满但是力量不足，成熟稳重但活力不够。因此，需要借助大数据，从根本上为其提供能力支持，这样企业才能够变得生机盎然，焕发出无穷的青春活力。

那么，如何才能让大数据与企业架构实现完美“结合”呢？为了实现这样和谐的“联姻”，还需要从企业架构和大数据两个方面分别做起。

对于企业架构，通过“分”与“合”，使得企业能够更好地完成战略目标。对于大数据，其特点和优势就是能够聚合全社会的“能量”，然后再通过抽取、转换、合并，最终形成满足企业需求的能力。

因此，企业架构与大数据的最佳结合点就是“能力”。企业架构从不同视角分解为不同的“能力”，而大数据则通过对数据的聚合形成了多种支持企业架构的“能力”，只有将企业架构的“能力”需求与大数据的“能力”供给对接起来，才能实现两者的完美“联姻”。

本章内容思维导图如下所示：



2.1 大数据与决策：选择远比努力更重要

分析后形成的决策决定了企业发展的方向与道路，影响深远，正确的决策会让企业靠近成功，而错误的决策必然会导致失败。

在人们的日常生活中，无论是个人还是家庭都会面对大大小小的决策：选择什么品牌的衣服，去哪家饭馆吃饭，乘坐什么交通工具，住哪家旅馆，等等，不一而足。

对于企业，每天同样需要面对各种各样的决策：客户的真正需求是什么，企业需要提供什么样的产品和服务？客户的服务等级多高，不同等级的客户可以享受到什么样的服务？销售人员的个人贡献多大，应当得到多少奖金？如何制定企业发展战略？等等。

对于提供公共服务的政府、事业单位、非营利组织，同样需要做出各种决策，无论是基础设施建设还是外交、金融等决策：例如，如何规划设计地铁路线？如何确定地铁票价？如何设定水、电、暖气的价格？未来五年国家在技术方面的主攻方向是什么？如何有效预防和控制艾滋病？等等。

对于企业特别是大中型企业来讲，企业关注的重点包括：如何提高管理水平，降低企业内耗，提高总体运营能力；在经济全球化、一体化的大背景下，企业如何快速响应外部市场的变化？

“适者生存、优胜劣汰”，在这样一个充满竞争的年代，企业只有顺势而为，采用先进的、科学的方法、技术与工具，才能在残酷的外部竞争中得以生存和发展，对于拥有国家垄断资源的大型国有企业，更应当积极变革，完成好国家交付的历史重任，降低服务价格，为老百姓带来更多的实惠。

当前，具有 4V 特征的大数据，可以通过分析历史预测未来，帮助组织发现商业与社会中存在的规律，进而指导组织进行决策，比如著名的奥巴马竞选案例、啤酒和尿布的故事，等等。

不同类型的组织均可以利用大数据服务。政府部门可以利用它分析社会舆论作为改革的参考依据，分析城镇化、老龄化等经济社会发展趋势，以完成政策的制定和合理的资源配置。企业则可以利用大数据实现对市场的调研，客户特征和行为分析，开发适合市场需

要的产品，实现精准化营销等。个人则可以利用大数据进行职业规划、确定出行计划、投资理财决策等。可见，大数据可以应用于生产生活的许多方面。

在大数据概念出现之前，人们已经在决策支持领域进行了长期的研究，形成了许多研究成果，包括数据仓库、数据挖掘、商业智能、在线分析等，在开始本节之前首先看一下大数据服务与数据挖掘等以往的概念的区别和联系。

同大数据服务一样，数据挖掘、商业智能等同样用于对组织决策的支持，数据挖掘领域的经典案例就是啤酒和尿布的故事，通过基于客户购买习惯的分析，发现客户在买尿布的同时也会购买啤酒，这一规律的发现可以帮助卖家确定如何摆放商品，即商家可以根据这一发现将啤酒和尿布摆放在一起，这样顾客就可以更加方便地拿到商品，既提高了客户满意度，也提高了商品的销售能力。

当然，在“数据挖掘”时代，社会的数据规模还没有那么大，数据存储通常还是关系型数据库，被处理的数据也多数是结构化数据，在大数据时代，数据的规模大大提高，采用传统的方法和技术已经难以实现，这是大数据不同于“数据挖掘”时代的一个特征。此外，“数据挖掘”时代更多地强调如何将数据分析的结果提供给决策者使用，在大数据时代，大数据服务的作用除了供决策者参考之外，更加强了决策的自动化，通过决策自动化提高组织效率并降低成本。

2.2　张开想象的翅膀：大数据服务畅想

技术是手段，业务发展才是最终目标，企业首先需要从战略、建设、产品、客户、供应商、人才物等业务视角畅想可能需要的大数据服务。

“只有想不到的，没有做不到的”，历史证明，只要是人类能够想象到的，迟早会变为现实。比如古代由于科学技术并不发达，人们只能在头脑中想象“嫦娥奔月”，在几千年后，人类掌握了飞船技术，终于将千年前的梦想变为现实。

本节从企业的业务活动出发，从战略、建设、产品、客户、供应商、人财物等多个方面对大数据服务进行了畅想，以免限制大数据服务的想象力。

2.2.1 大数据与战略管理

企业发展战略的目标是实现“知己知彼，百战不殆”。“知己”是要了解企业自己的特点，有什么优劣势，企业自身拥有多少人、财、物等资源。“知彼”则是要了解企业的竞争对手情况以及企业所处的外部环境，包括地域政治环境、经济发展水平、社会文化生活状况、行业环境、区域环境、技术发展趋势等。

相对于“知彼”，“知己”要容易得多，就像人们谈论自己容易一些一样。因此，下面首先谈一谈企业如何借助大数据实现“知己”。

对于已经实施信息化的企业，借助传感器和信息系统实现了对企业生产经营活动的支持，同时也记录了企业的各种业务活动。比如企业的客户关系管理系统中记录了客户、产品、营销、销售、订单、维系、挽留、服务等信息，呼叫中心平台中记录了客户咨询、投诉、申告、建议、通话录音等信息，计费账务系统中记录了客户账单、详单、充值、缴费等信息。业务部门人员产生的数据如图 2-2-1 所示。

除了业务信息的记录，企业通常还会通过人力资源管理系统、财务管理系统、资产管理系统等记录人力资源、财务、资产等信息。例如，人力资源管理系统记录了员工姓名、年龄、教育经历、工作经历、培训经历、入职时间、技能特长、所在岗位、工资、津贴等信息；财务管理系统记录了历史投资、融资、资产负债、损益、现金流、预算、成本等信息，资产管理系统则记录了资产名称、资产原值、折旧年限、折旧方法、资产现值、归属部门、责任人等信息。职能部门人员产生的数据如图 2-2-2 所示。

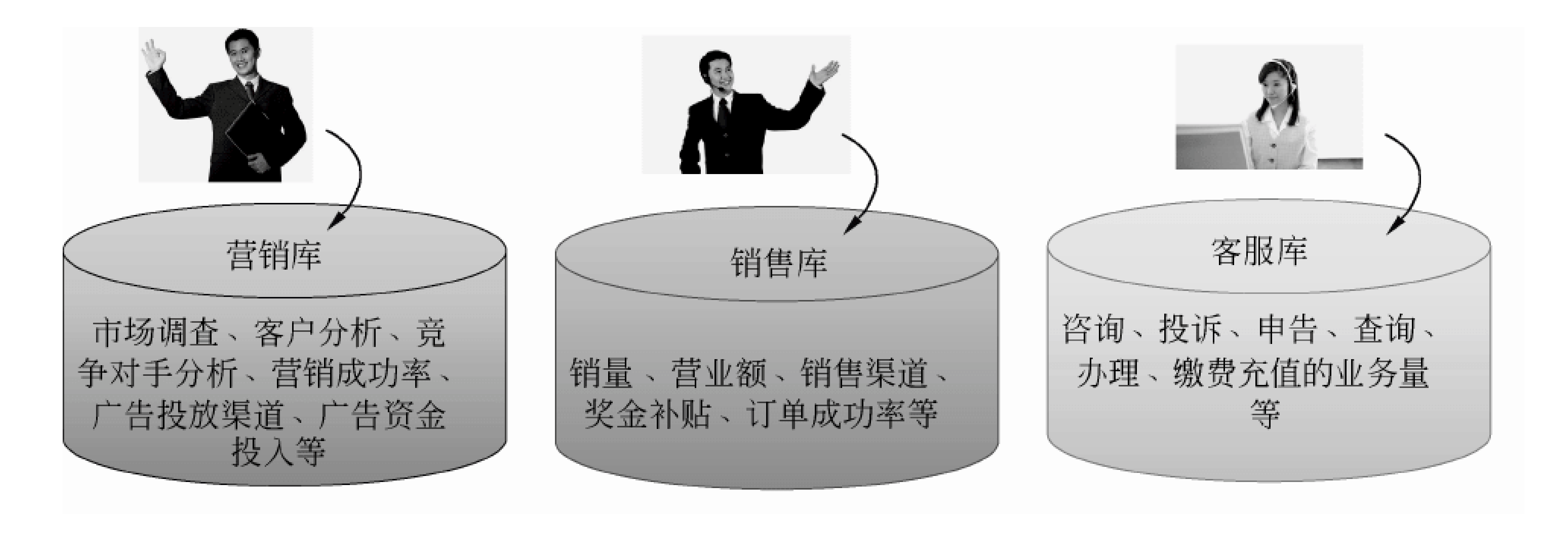


图 2-2-1　业务部门人员产生的数据

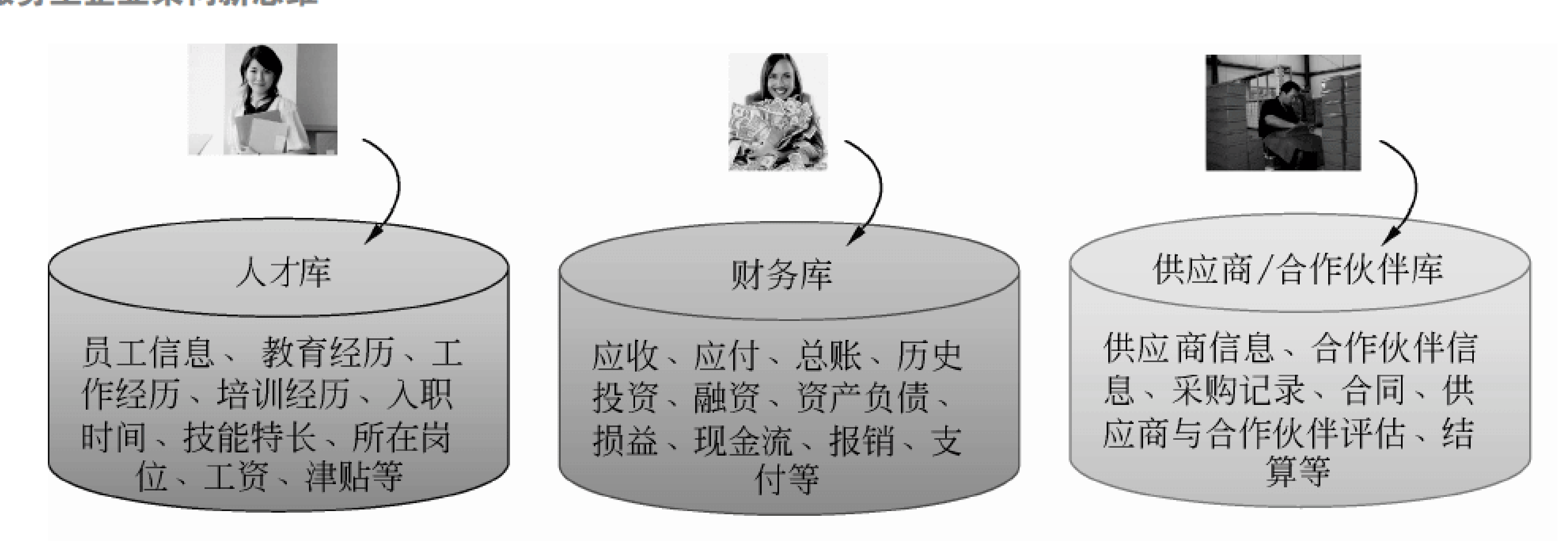


图 2-2-2　职能部门人员产生的数据

由于以上信息和数据属于企业所有，因此比较容易获取。可见，对于企业战略管理人员来说，应当优先考虑对其进行收集和分析，以实现“知己”。

与“知己”相比，实现“知彼”的难度相对大一些。首先，由于企业之间竞争的原因，企业的竞争对手会设法保护自身的信息和数据，以免在竞争中失去优势，比如苹果公司就是在新产品发布之前，对外界屏蔽新产品相关信息，以免竞争对手对创意进行模仿，妨碍产品的营销推广；其次，企业的外部发展环境信息往往受到多种因素的干扰，比如信息化水平、数据开放水平等。一般来说，社会的数据开放水平越高，企业对于宏观发展环境的掌握越准确，越能够制定正确的发展战略。

企业实现“知己”主要是通过采集企业自身的信息系统获得的，而要实现“知彼”则需要借助一定的方法和手段，从多样的渠道获取。比如，行业发展动态可以来自于专业部门的网站，比如，电信用户规模、增量、增速等数据可以从工信部网站获取；人口结构、收入水平、消费能力、法律法规等数据可以从国家统计局网站获取。社会习俗、文化传统等则可以从图书馆管理系统中检索获取。此外，报纸杂志、咨询公司研究报告、知名网站也是情报获取的重要数据源。

数据获取的方法和手段也是多种多样的，有效的、真实的数据获取往往需要采用非常规手段。据悉，某大型跨国外企甚至从竞争对手的垃圾桶中获取情报，这与战争时期利用特务机构获取情报的道理是类似的。

总之，企业获取的情报越准确、越及时，越能够帮助战略制定者制定正确的决策。由于企业发展战略决定了企业发展的方向和道路，决定了企业的未来，因而其重要性不言而喻，在互联网时代，大数据对于企业战略管理将变得越来越重要。

2.2.2　大数据与建设管理

基础设施是支撑企业运营的物质基础，是企业战略落地的第一支撑点，也是企业产品设计的前提条件。

企业需要按照全生命周期的管理方法来管理基础设施，包括基础设施的规划、设计、采购、库存、验收、上架、运行、下架、退出的全过程。基础设施生命周期管理过程如图 2-2-3 所示。

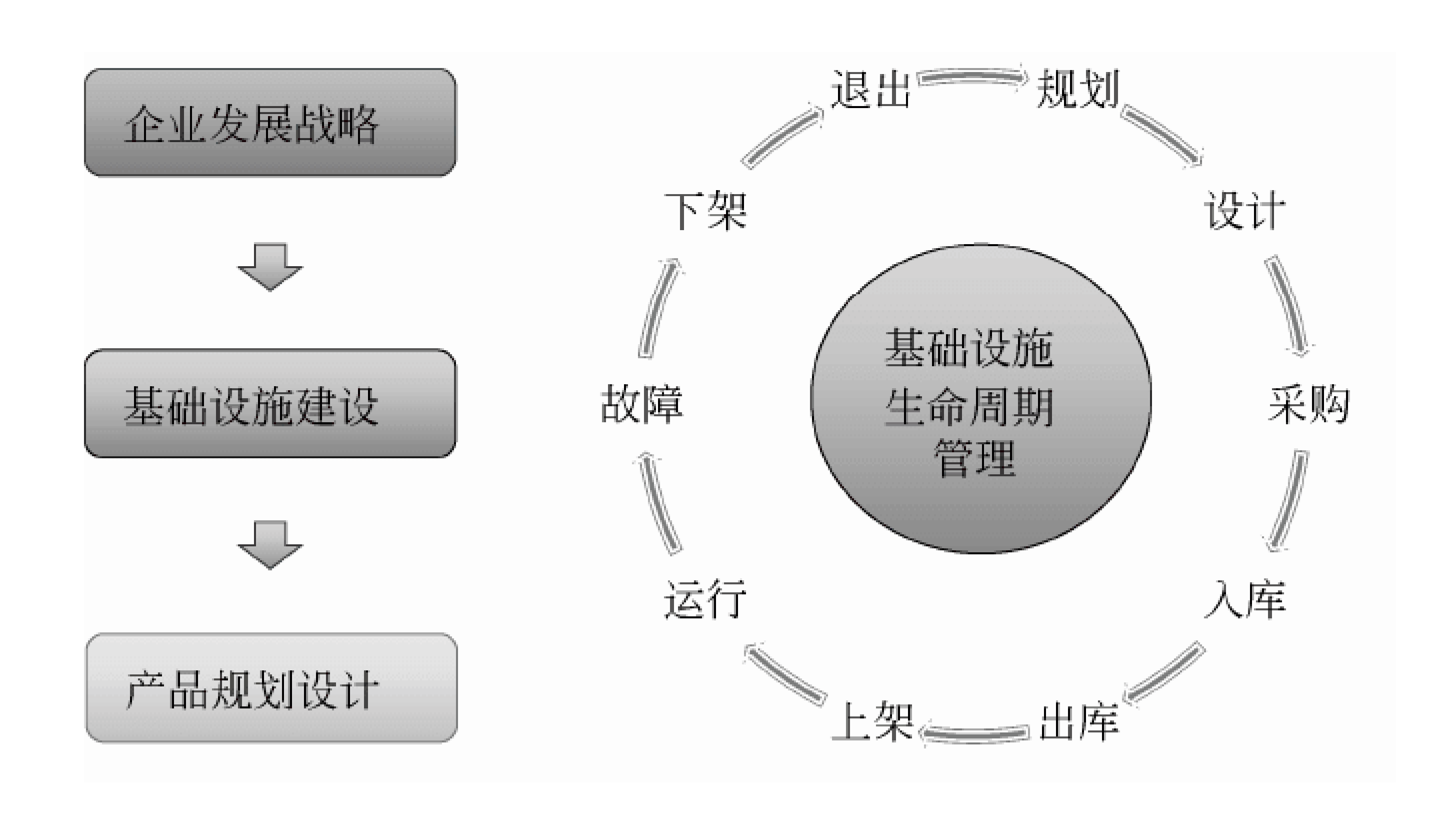


图 2-2-3　基础设施生命周期管理

大数据对于企业基础设施全生命周期的各个阶段都具有重要的作用。在基础设施的规划设计阶段，可以借助大数据来计算基础设施建设的位置、规模、容量等；在基础设施采购阶段，可以利用企业内部、采购网站等不同渠道对该产品或类似产品的评价等大数据判断基础设施的性价比，为采购决策提供参考；在基础设施的运行阶段，可以借助采集到的运行数据（平均故障时间、负荷指数、性能等）作为基础设施扩容或退出的参考依据。

以电信运营商为例，在网络的规划设计阶段，可以基于用户位置、移动宽带业务使用、用户价值高低、用户业务访问网络路径等大数据作为某个地域的网络投资建设或者扩容的依据；在网络设备的采购阶段，可以收集同类网络设备的历史运行数据和来自于互联网渠道的用户评价等来判断是否采购该网络设备以及采购价格区间等；在网络设备的运行阶段，可以通过采集网络设备的历史运行数据并进行分析，来决定该网络设备是否需要扩容或者报废。

大数据在通信网络设施全生命周期的作用如图 2-2-4 所示。

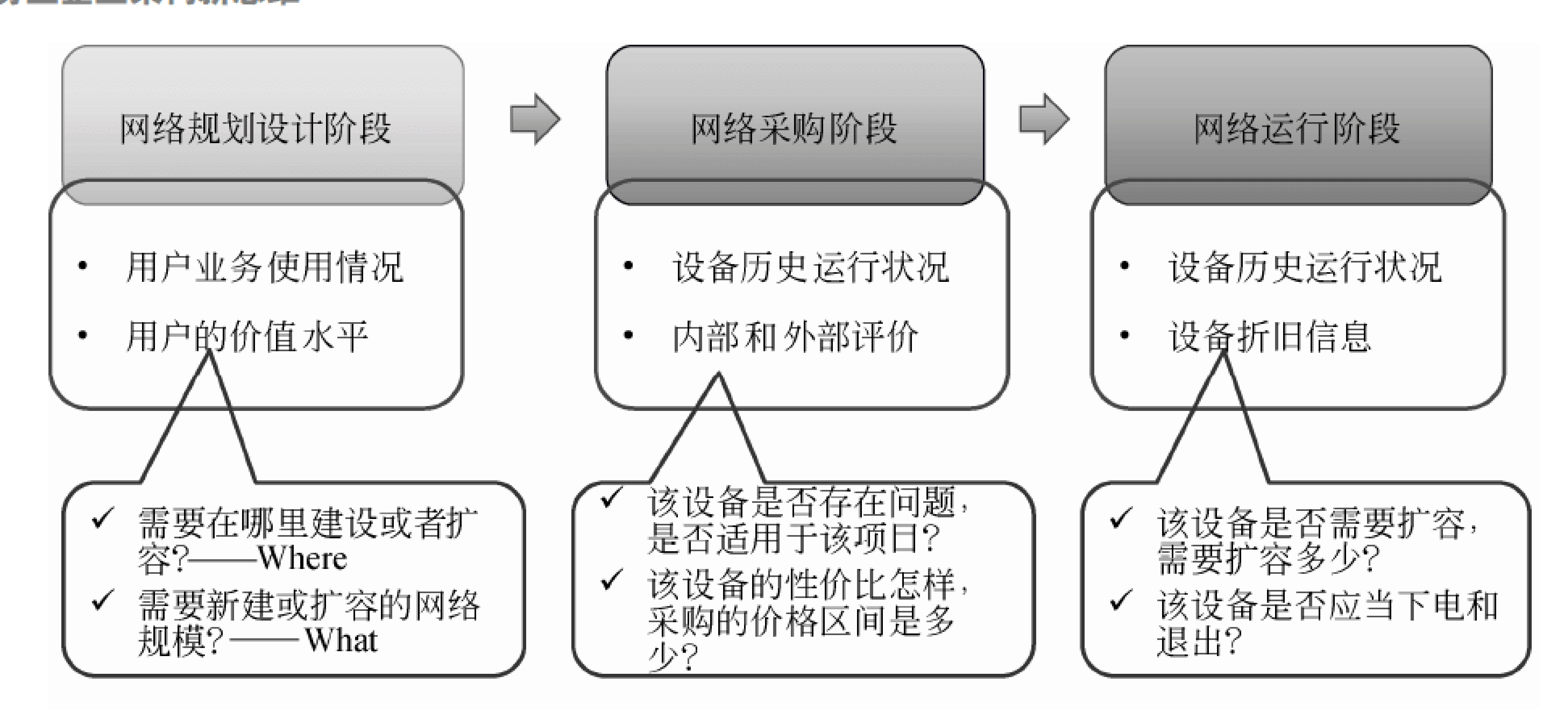


图 2-2-4　大数据在通信网络全生命周期的作用

从上图可以看出，大数据在通信网络全生命周期中的各个阶段都能够发挥作用。之所以能够做到这一点，是因为通过对不同来源的大数据进行汇聚、整合与分析，形成了作为基础设施建设的客观依据。

比如，在网络建设阶段，企业可以借助用户的业务使用、用户的价值等大数据，分析并得出是否需要新建或者扩容网络以及新建或者扩容规模的预测。

同样的道理，在网络设备采购阶段也可以借助同类设备的用户评价、不同供应商的价格对比等，分析并得出该设备是否存在质量问题，是否适用于该工程项目，该设备在什么价格区间才能够保证较高的采购性价比等。

在网络的运行阶段，可以借助设备历史运行状况、设备折旧信息等来判断该设备是否还能够满足当前应用的需要，如果不满足则应当扩容多少，该设备是否已经到了报废年限，是否应当下电等。

总之，通过汇聚并整合来自不同数据源（使用、评价、运行、折旧等）的数据形成的大数据服务，可以有效地支持企业对基础设施全生命周期的管理。

核心资源通常是界定行业界限的关键，比如通信行业的核心资源是通信网资源，金融行业的核心资源是货币资源，互联网行业的核心资源是信息资源。对于服务型企业，资源则主要体现在对质量的保证（QA）方面。

通信网络资源是电信运营商最核心的资源。从专业划分的角度看，通信网络资源分为传输、交换、管道杆路、无线接入、固网接入、平台、IT、局房、电源、电气等多个专业；从网络的拓扑结构看，通信网络资源分为核心网和接入网，接入网又分为有线接入网和无

线接入网；从网络的层次结构看，通信网络资源又分为本地网、省干网（二干）、骨干网（一干）以及跨越国界的国际网。

从通信网络资源的生命周期角度看，包括规划设计、工程建设、运行维护、下架退出四大阶段。那么，大数据价值就是支持通信网络资源全生命周期的管理。如图 2-2-5 所示。

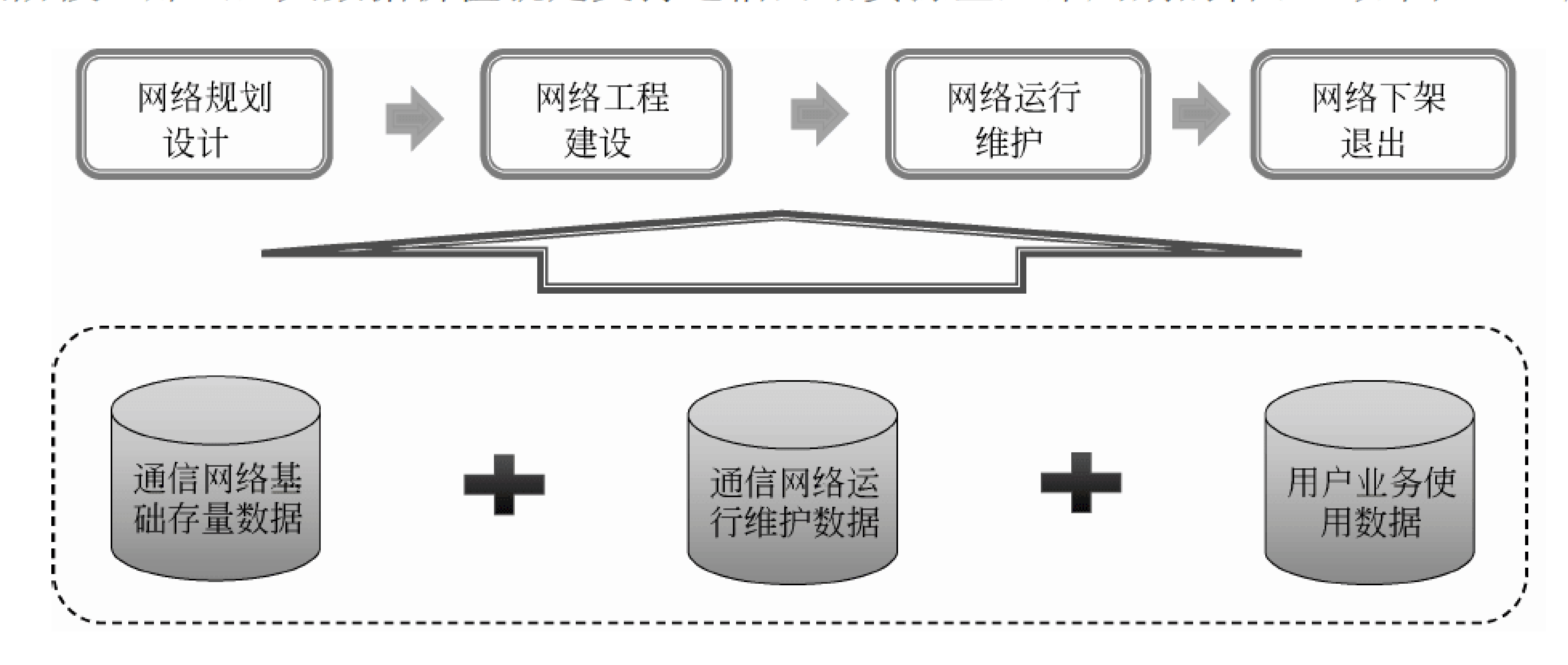


图 2-2-5　大数据对网络资源管理全生命周期的支持

在规划设计阶段，可以利用来自通信网络基础数据、运行数据以及用户使用数据，为通信网络的规划设计提供数据基础。以 4G 网络规划为例，首先是收集所有基站的基础数据，然后再根据移动用户上网记录来生成基站的流量数据，最后根据基站流量数据来判断基站扩容的可行性以及扩充的容量大小，为基站建设和扩容提供参考依据。

在工程建设阶段，可以借助大数据辅助完成通信网络产品采购与验收。对于通信网络资源的采购决策，最有说服力的莫过于对设备运行效果的评价。通信网络设备采购的主要考虑因素包括可靠性、可扩展性、可用性、性能、安全性、价格等。对于可靠性，可以通过待购设备类似产品的运行结果和测试报告来验证，其他方面则可以通过测试后获取的数据来验证。可以对来自多个供应商的设备配置、价格等进行综合分析，计算出设备的性价比。同时，还要将企业注册资金、资质认证、行业应用案例等因素综合起来考虑。

在网络运行维护阶段，可以借助大数据进行问题根源分析，形成知识库。网络管理系统可以对网络设备的运行状况进行监控，根据设备的运行效果来评价设备的质量指标。

2.2.3　大数据与产品管理

产品之于企业，对外满足客户需求，对内则体现为资源占用，是连接企业内部和外部

的桥梁和纽带。一方面，企业通过为客户提供产品而获得收入，另一方面，产品的供给则体现为企业内部资源的消耗。外部用户产品使用和企业内部资源消耗之间的关系如图 2-2-6 所示。

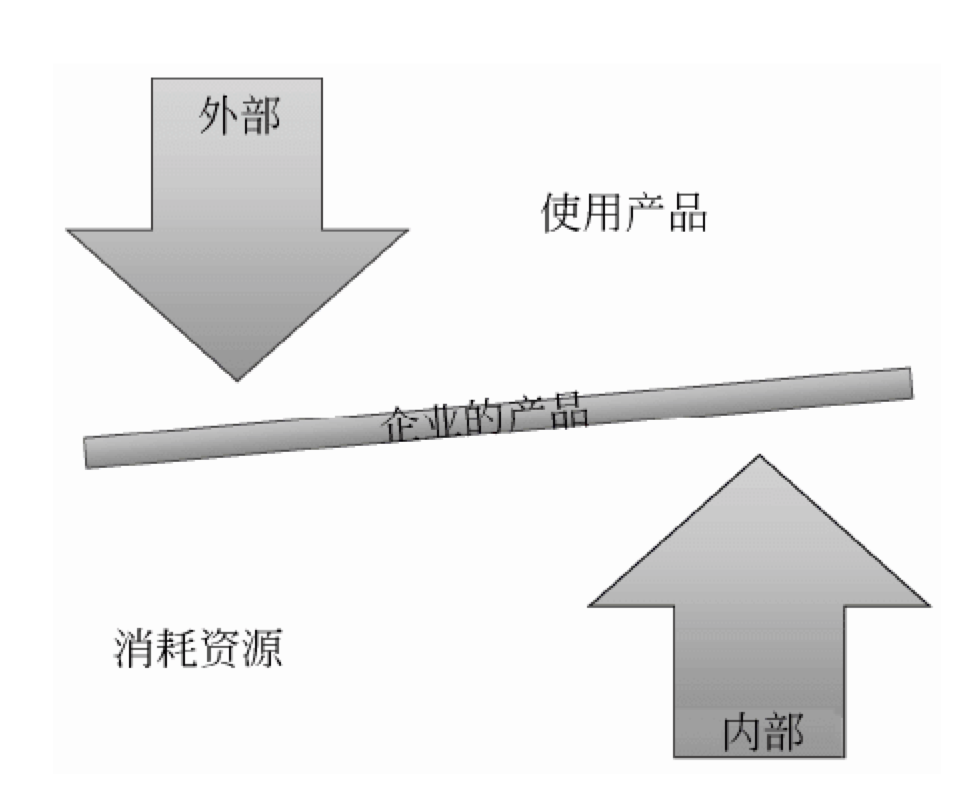


图 2-2-6　外部用户产品使用和企业内部资源消耗之间的关系

为了方便分析，将产品的功能分为收入产生和成本消耗两个方面。

企业收入产生的过程，其实就是产品销售的过程。通过产品的销售，企业可以掌握产品的特点，包括购买人群特征、销售规模、销售额度、不同时段的销售情况等，企业可以将这些数据作为判断产品价值高低的依据，如果分析发现该产品为市场活跃型产品，则可以继续推广。

企业产品成本消耗的过程，其实就是资源消耗过程。企业的产品需要在消耗各种资源后形成，消耗的资源包括人工、机器设备、材料等。相对于收入产生而言，成本的消耗则要复杂得多。

如果对企业的产品带来的收入与其消耗的成本做个减法，就算出了企业的利润。如果为正就说明该产品能够使企业盈利，为负则说明企业亏损。当然，这只是产品评估的一种基本方法，企业有时会根据企业整体发展战略来调整产品市场战略，比如对有些产品采用免费或者低价策略，以便迅速占领市场。

与基础设施生命周期一样，产品也要经历定义、开发、导入到成长、成熟、衰退的过程。大数据可以应用于企业产品生命周期的不同阶段，帮助企业进行产品的定义、开发、导入以及性能优化。产品生命周期及其分析方法如图 2-2-7 所示。

内部成本效益分析包括产品销售数据获取、产品成本数据获取、产品成本效益分析，分析结果可以作为产品定义或者退出的参考依据。

外部环境分析包括客户对产品的评价数据获取、产品相关新技术信息获取以及竞争对手的同类产品信息获取。分析结果可以用于产品的性能优化、产品的定义、产品退出的参考依据。

以通信产品为例，通过对来自社交网站评论数据的分析发现，许多人抱怨给家里的老人或者孩子的手机经常性地分开充值比较麻烦，希望能够只给一个号码充值，其他号码就可以共享这个号码的余额、时长、流量等。企业通过对来自社交网站数据的进一步分析发现，这些客户通常是工作比较忙碌的中年人，同时他们通常也是高价值客户。为此，电信运营商定义和开发了一种可以实现多个号码捆绑并且号码可以直接共享余额、通话时长、流量等的新型产品，从而节省了该类用户的充值缴费时间，减少了充值的次数，为该类用户带来了更大的便利。

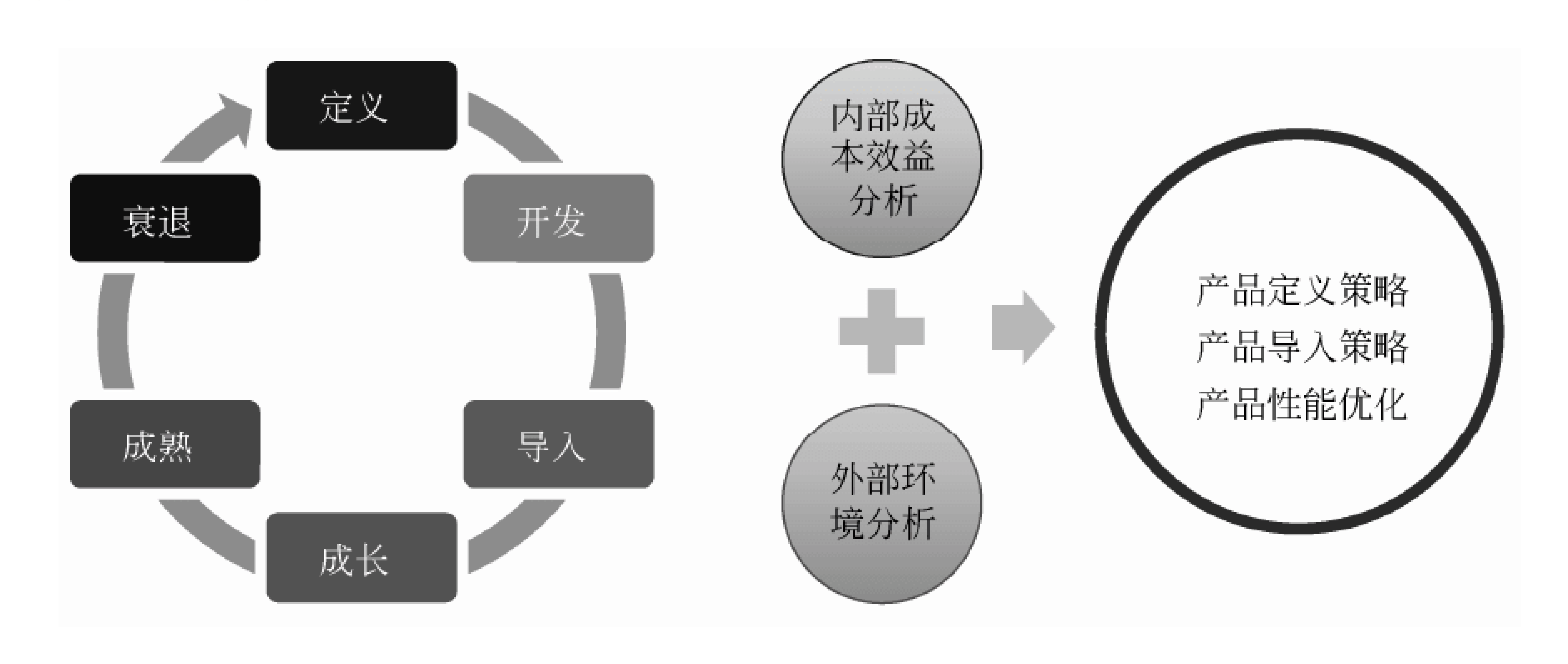


图 2-2-7　产品生命周期及其分析方法

产品定义和开发的需求也可能来自于企业的竞争对手。某电信运营商通过对来自社交媒体的数据分析发现，许多客户对竞争对手的定制化产品评价很高，原因是竞争对手的产品可以实现业务的定制化销售，客户可以自行选择语音、上网、短信业务的月使用量，同时竞争对手还会根据客户的历史消费情况为客户推荐一款最适合客户的套餐。这种方式提高了客户选择产品的灵活性，降低了因为客户选择不合适的套餐而产生的不必要支出，为客户节约了通信费用，同时也提升了竞争对手在客户心目中的品牌形象。于是，某电信运营商也借鉴竞争对手的产品管理经验，快速地推出了可定制的、具有套餐推荐功能的产品。

再以金融行业的金融租赁产品为例，金融产品的特征主要体现在风险控制上，而风险系数的高低主要取决于对于承租方和租赁物的了解。通常来说，高风险预示着高回报，因此金融租赁公司的业务发展主要取决于对风险的控制能力。由于高科技产品更新换代快，

也就意味着因为技术的发展变化而带来的较大的投资风险。这时，金融租赁公司可以借助大数据，收集关于行业、产品以及与该产品有关技术的发展情况，科学地评估租赁物的风险。以目前提供云服务的数据中心业务为例，随着数据中心业务的快速发展，数据中心的可靠性和稳定性也变得越来越重要，保障可靠性和稳定性的关键产品就是 UPS，金融租赁公司可以与 UPS 厂商合作，推出相应的厂商租赁产品。

再来看一看大数据对于互联网行业的影响。近年来，以开放、合作为特征的互联网飞速发展，由于互联网产品进入市场的门槛低，在激烈的市场竞争下，许多产品被淘汰，产品生命周期很短，因此互联网产品更加讲究以客户为中心进行设计。以阿里巴巴集团的支付宝为例，通过市场调研发现，随着电子商务的发展，网民经常购买小额商品，而银行卡支付存在支付不方便、安全性差等问题，余额宝可以同时关联多张银行卡，用户可以预先从银行卡中转入小数目的资金，这样用户就不用担心支付安全问题，同时用户只需输入支付密码即可完成支付，大大提高了支付的便捷性，满足了人们高频率的网上购物需求。

可见，大数据之所以能够为产品生命周期过程提供帮助，主要是企业利用大数据可以发现市场规律和客户需求，并根据市场需求提供满足客户需求的产品。大数据为企业准确地了解市场需求创造了条件。

2.2.4 大数据与客户关系管理

在卖方市场中，由于产品稀缺，因此企业是上帝，客户为了买到商品，往往需要通过找关系甚至请客送礼才行，这是我国实行计划经济年代的现象。自从我国全面发展市场经济以后，社会商品大大增多，市场变成了买方市场，客户可以根据自己的喜好在多家企业的商品中进行选择，企业为了销售自己的产品，反而需要与客户搞好关系，否则就难以获得利润，甚至破产倒闭。可见，在买方市场的情况下，客户关系对于企业是多么重要。

从企业角度看，客户关系通常要经过建立/恢复、维持、挽留、终止四个阶段。企业视角的客户生命周期如图 2-2-8 所示。

在客户关系建立阶段，企业通过广告宣传等手段来吸引客户购买企业的产品，如果为企业曾经流失的客户，则根据企业与该客户以往的接触记录采取更有针对性的营销行动。在客户关系维持阶段，企业通过积分送礼品、寄送生日礼物等方式回馈客户，增进客户对企业的感情。在客户关系挽留阶段，企业通过分析客户离开的原因有针对性地采取补救措施，比如给予新的产品折扣、推荐新产品、延长服务期限等。

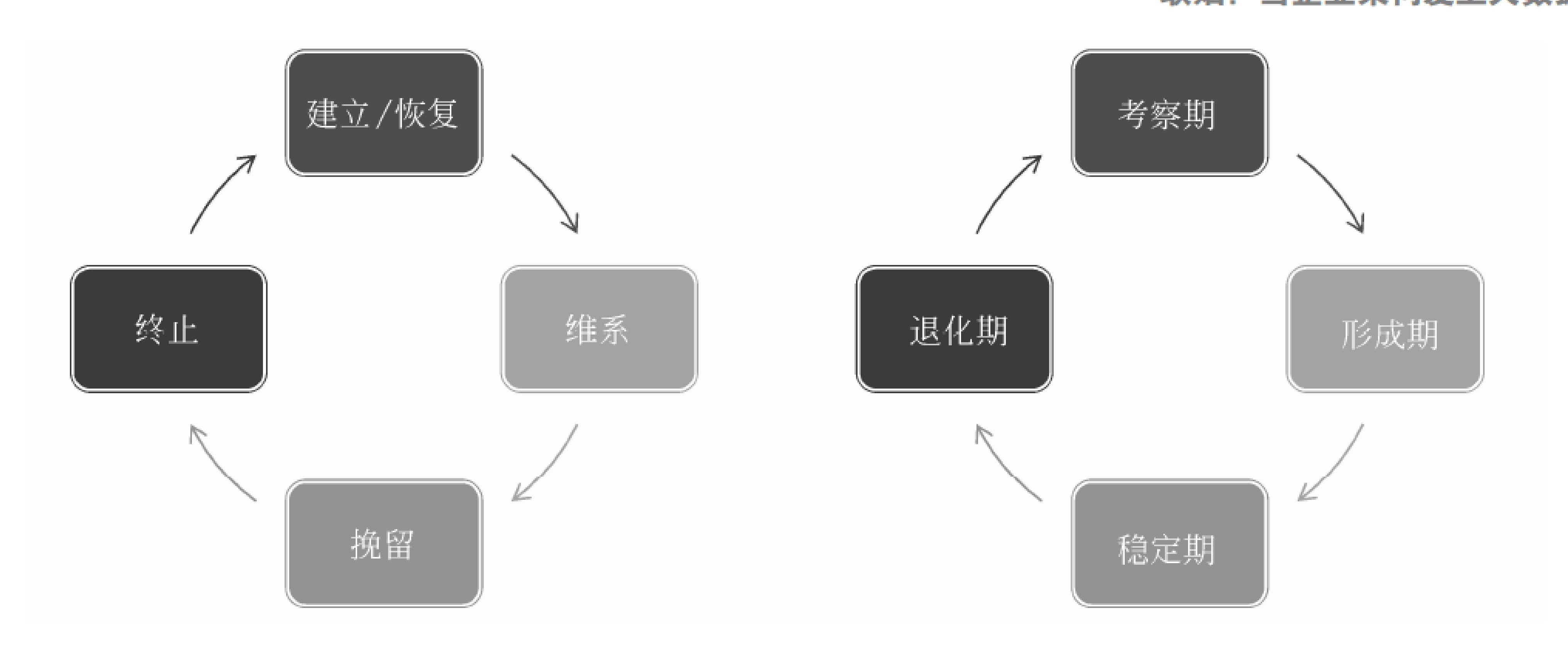


图 2-2-8　企业视角的客户生命周期

在客户对企业的产品和服务进行咨询、购买、使用、付费、申告、投诉、建议的过程中，企业借助信息系统记录了客户的信息，包括客户的属性信息和客户的行为信息。以通信产品为例，客户属性和客户行为如图 2-2-9 所示。

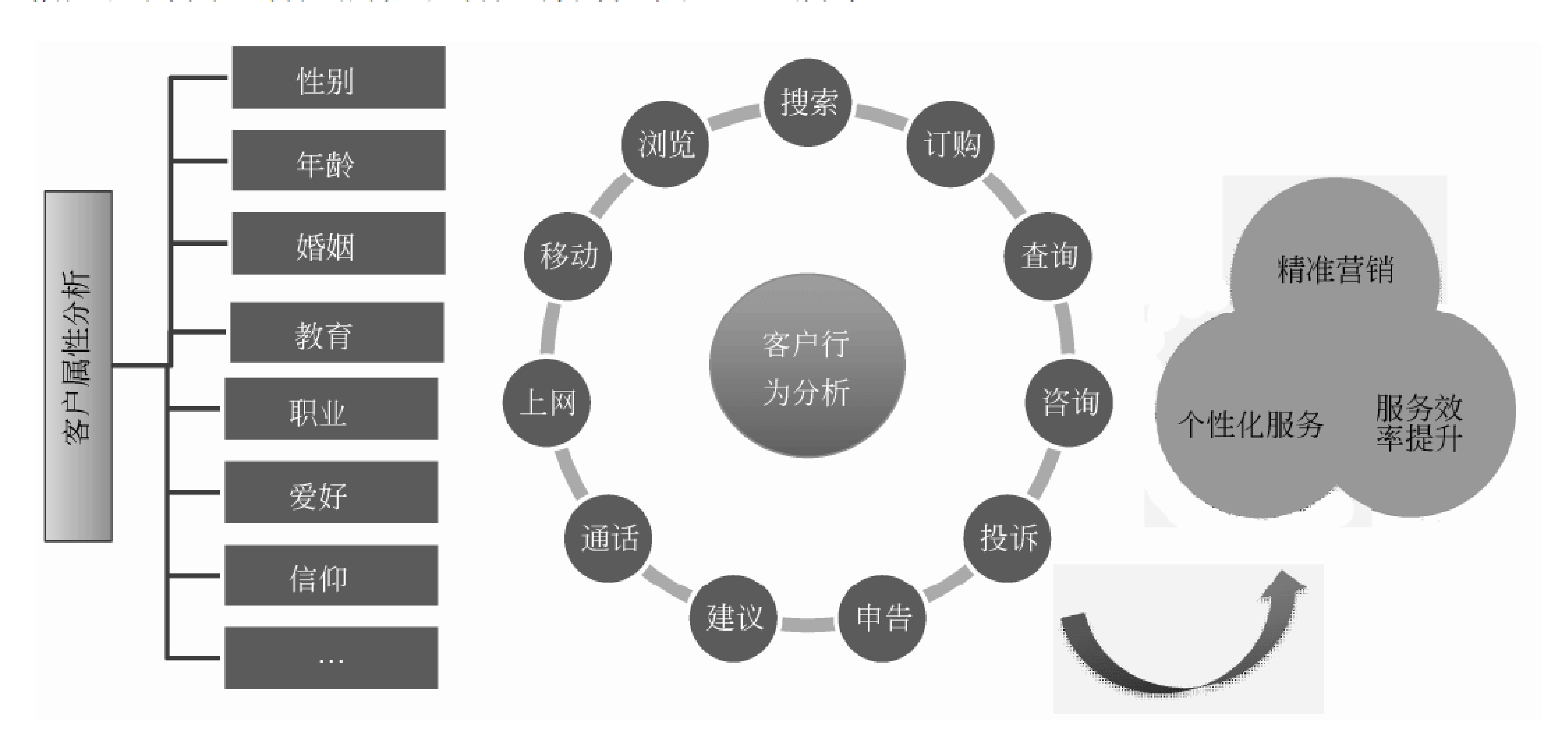


图 2-2-9　通信企业客户特征与行为画像

企业为了建立与客户的关系，需要两类数据源作为支持。第一类数据是来自社会大环境的数据，包括人口、经济、政治、法律、文化等方面，获取这些数据的目的是帮助企业找到目标市场和目标客户群。比如根据某地域人口的年龄结构分析，发现该地区老年人的人口比重越来越高，因此企业可以向该地区推出适合老年人的产品和服务。第二类数据来

自曾经使用过企业的产品或服务，但是当前已经流失的客户，企业可以非常容易地获取这类客户的信息，企业应当尽力基于客户以往的行为来发现客户的偏好，重建与客户的关系。

以金融行业的银行为例，在客户关系建立阶段其主要业务活动是对客户进行信用评估，以便确定客户的风险敞口。如果为个人客户，则需要客户的收入、学历、工作单位类型等作为确定信用额度的依据。如果为企业客户，则需要收集企业客户的经营情况和财务情况。分析企业经营情况的目的是判断企业产品是否具备市场竞争力、企业是否具有持续的盈利能力，通过分析其盈利能力来判断其偿债能力。分析企业财务情况的目的是判断企业是否存在财务风险，评估的数据来源包括资产负债表、损益表（利润表）、现金流量表等。

企业为了维持与客户的关系，需要两类数据源作为支持。

第一类是来自客户对企业产品或服务的故障申告、投诉、建议等，通过这些数据可以发现客户在产品或者服务使用过程中存在的问题，并尽快解决产品中存在的问题，提升客户服务水平。

第二类来自企业对客户产品或者服务使用过程中存在问题的主动发现，在客户还没有进行故障申告或者投诉的时候就修复问题或者主动对客户进行提示，增强客户对企业的好感。比如电信运营商可以使用移动用户的上网记录来主动发现客户上网是否存在速度问题，是否存在基站容量不足等问题，如果通过分析发现客户访问某些应用的数据量大并且访问该应用的客户总体 ARPU（单用户平均收入）值高，说明该应用以及访问该应用的客户群体都是高价值的，同时发现移动用户到该应用之间是跨电信运营商网络的，由于不在同一个网络，根据经验判断该区域的移动用户访问这个高价值应用的速率一定不高，因此可以建议该应用的提供商在该区域增加 CDN（内容交付网络）节点，解决该区域移动用户对该应用的跨网访问问题，通过缩短移动用户到应用之间的网络访问路径来提高移动用户上网体验。

当客户具有离开企业的倾向时，企业应当及时发现并赢回客户。引起客户离开企业的原因包括：客户自身原因、企业原因、竞争对手原因。企业应当利用大数据来预测客户离开的真正原因。如果是客户自身的原因，比如客户喜欢尝试新的产品因此换成竞争对手的产品，如果确认确为客户自身不可逆转的原因，企业不必在这类客户身上耗费太多的成本。如果是企业自身的原因，企业应当分析是哪一类原因，然后采取相应的补救措施。比如客户是因为企业的客户服务水平低而离开的，这时企业需要提高服务质量，如果客户是因为企业的产品质量存在问题而离开的，则企业需要提升产品质量。如果客户是因为竞争对手

原因离开的，则企业需要找出企业与竞争对手的差距，然后在这些差距方面进行提升。比如竞争对手推出了上网速度更快更稳定的产品，那么企业就应当采用先进的技术来提升产品质量。

那么，对于客户流失预测需要的数据源包括客户对企业的产品使用数据、客户服务数据、客户账户数据、企业竞争对手数据等。通过对以上数据进行整合，可以发现客户对产品的使用频率、客户账户余额、客户投诉内容、竞争对手产品资费对比等，进而判断客户是否具有离开企业的倾向。

互联网的发展，为人类提供了实体空间之外的另一个虚拟空间，借助这个没有边际界限的互联网，人与人之间的社交方式也发生了巨大的变化。新型的基于互联网的社交方式一方面可以从各种社交圈子中获取信息，另一方面个人也可以快速发布自己的见解，使得言论通过圈子迅速传播。在这个信息获取和信息发布的过程中，形成了大量的数据，为了保证信息传播的速度，同样需要大数据技术来作为支撑。

2.2.5　大数据与供应商/渠道商关系管理

随着经济发展的全球化和一体化，商业模式从单一链条的价值链（Value Chain）模式发展到网络模式，世界进入价值网络（Value Network）时代。在价值网络时代，企业之间的协同更加紧密，竞合关系更加复杂，企业需要更加敏捷地应对外部环境变化。

作为为企业提供输入的供应商，在价值网络时代变得更加重要，供应商产品和服务的提供速度、质量、价格等对于企业来说越来越重要。

另一方面，作为为企业交付产品的渠道商（分销商、零售商等），借助其商品整合能力和本地客户资源优势，在价值网络时代也成为企业重要的合作伙伴。企业与供应商、客户、渠道商等合作伙伴的关系如图 2-2-10 所示。

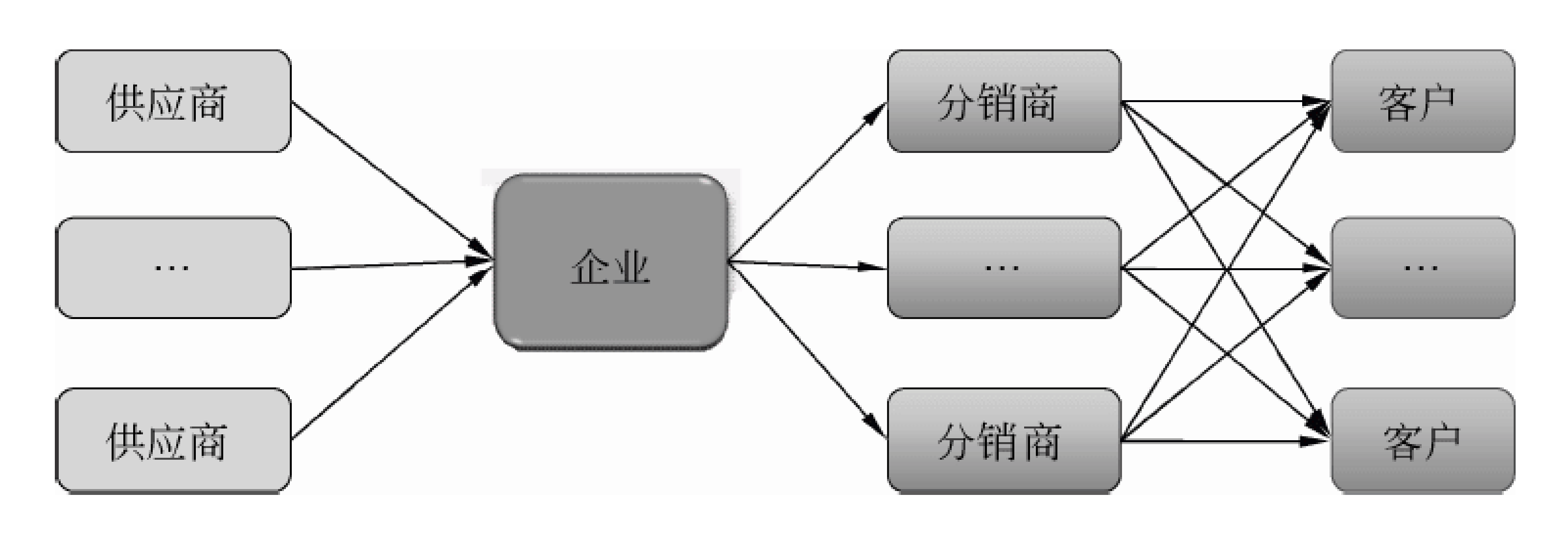


图 2-2-10　价值网络时代企业对外合作关系

每个企业都有提高产品和服务的供应商，同时也有帮助企业销售产品和服务的渠道商，下面就对大数据与企业这两种角色分别进行分析。

1. 大数据与供应商关系管理

供应商作为企业的输入部分，对于企业对外提供产品和服务起着重要的作用，尤其是企业的采购范围扩大到全球，加大了采购风险，供应商关系管理更加重要。

在大数据时代，企业应当实时地监控供应链运行情况并对可能存在的风险进行评估，迅速果断地采取补救措施。企业还需要打破企业内部采购、运营、营销等部门的竖井模式，实现内部信息共享。同样，企业还应当实现与外部供应商的信息共享，让供应商能够获知企业的市场情况，包括企业的客户对哪些产品感兴趣、对于产品的使用评价等，辅助供应商及时调整产品和服务的生产。

与客户关系管理类似，作为企业“输入”的供应商，同样需要进行关系管理，原因如下：首先，供应商为企业供应的产品和服务对于企业形成产品非常重要，如果企业没有好的供应品，那么企业也难以为其客户提供好的产品和服务，可见供应商提供的产品质量是非常重要的。其次，为了应对市场需要，企业需要供应商按照时间要求供应产品和服务，以便在市场竞争中争得先机。对于企业来说，对供应商的要求主要包括速度、质量、价格三个方面。速度决定了企业向客户提供产品和服务的速度，而质量则决定了企业向客户提供的产品和服务的质量。价格则决定了企业向客户提供的产品和服务的价格高低。企业应当综合平衡产品获取速度、质量以及价格。

从大数据对于供应商关系管理的支持角度看，企业需要借助大数据全面、准确地获取所需供应品的价格、质量、供应商信誉等信息，寻找高性价比的供应品，为了保证企业在市场中的竞争力，需要与供应商建立战略性伙伴关系，以保证供应品的稳定交付。不同渠道的供应商数据与供应商关系管理目标如图 2-2-11 所示。

企业为了实现对供应商关系的有效管理，提升自身产品的市场竞争力，需要从尽可能多的渠道收集供应商相关数据，包括供应商的产品、价格、运营、财务、资质、信誉等数据，以便及时掌握各供应商的产品情况，降低生产经营风险。

2. 大数据与渠道商关系管理

广义上讲，合作伙伴涵盖一切与企业有合作关系的组织，包括设备供应商、内容提供商、服务提供商、分销商、零售商等。为了区分合作伙伴与供应商的关系，本书中的合作

伙伴特指渠道商。渠道商包括分销商、批发商、零售商、代理商等一切帮助企业销售产品和服务的商家。

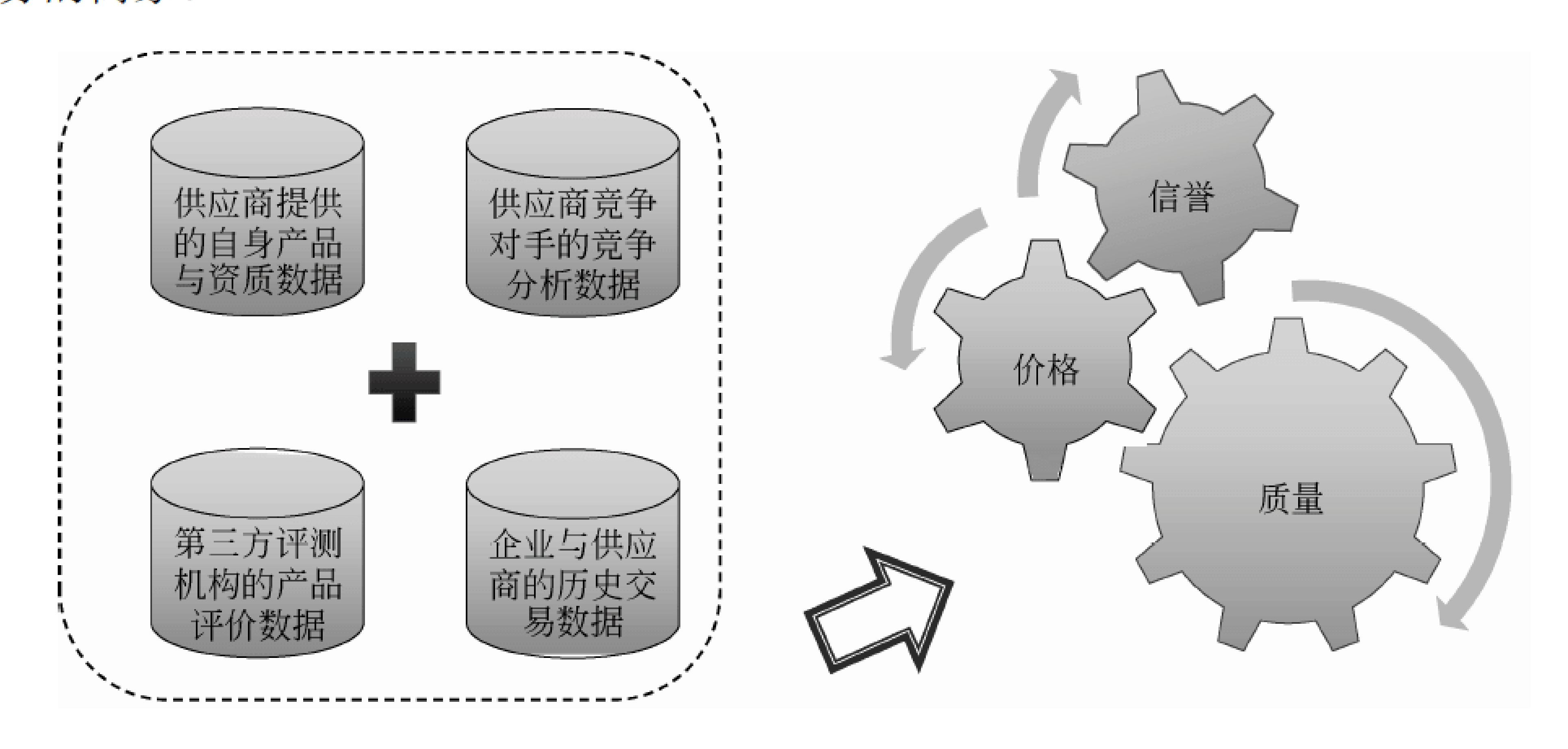


图 2-2-11　与供应商相关的数据与企业的关系

渠道商整合了来自多个商家的商品，因此渠道商提供的商品齐全，种类丰富，为客户提供了更加便捷的商品和服务，由于渠道商直接掌控大量的客户资源，因此具备强的议价能力，可以提供价格更低的商品。因而，在商品极其丰富的时代，体现了渠道为王的特征。例如，沃尔玛、家乐福等主打家庭生活用品的销售，而苏宁电器、大中电器等则聚焦在家电产品的销售，成为当代成功的渠道商。

对于企业来说，只有做好与渠道商的关系管理，才能够帮助企业将产品和服务传递到客户手中。当然，企业与再售商合作中提高销售能力的同时，也会以牺牲销售佣金为代价。与企业与供应商的关系管理类似，企业与渠道商的关系管理同样包括渠道商信息管理、准入管理、合同管理、佣金管理等。企业也应当积极寻找最适合的渠道商，设置准入门槛，建立和健全退出机制。此外，企业还需要建立针对渠道商的激励机制。

与企业与供应商的关系管理类似，企业与渠道商之间同样需要建立完善的信息共享机制，企业应当通过采集来自渠道商的产品销售数据和客户服务数据，发现产品的销售情况、客户对于产品的评价等，以便确定产品规划重点并改进产品质量。

2.2.6　大数据与计费收费管理

企业为客户提供产品或者服务，相应的客户也需要按照合同约定为企业支付产品或者

服务的使用费用。企业向客户收费的方式有多种类型，比如时长、次数、流量、利息、租息等，时长是一种按照使用期限收费的方式，多用于服务型企业，包括年、月、日、时、分、秒等多种方式。对于用于销售的产品，由于让渡了产品的所有权，因此往往是一次性收费，对于无形的服务，通常基于客户使用来进行收费。比如，电信运营商提供的通信服务，通常根据用户的通话时长、上网流量、发送次数等进行收费。对于提供存贷款业务的银行，往往根据资金占用的时间成本来收费。对于互联网提供的信息服务，通常采用对用户免费对投放广告的企业收费的反向收费模式。

以电信运营商为例，其计费过程要经过使用记录采集、合并、格式化、分拣、去重、批价、优惠、出账的过程。随着市场竞争的日益激烈，为客户提供实时消费情况查询的功能越来越迫切，而此时用户产生的上网行为记录越来越大，为了解决这一问题，迫切需要采用大数据技术。

比如像银行这样的金融机构，由于互联网金融的发展，人们可以通过多种第三方支付方式（比如支付宝、财付通等）完成资金支付，同时也需要从银行系统转账到第三方支付账户，因此形成了大量的银行之间转账的记录。为了提高用户对于实时交易记录的查询需求，也需要借助大数据技术实现高性能的查询。

大数据时代，数据规模大对于系统性能提出了越来越大的挑战，因此，如何实时地计算客户业务使用产生的费用，防止收入流失，成为企业控制风险的重要内容，因此，企业可以利用大数据技术，预测用户使用行为，提高海量数据的实时计费能力，及时进行风险预警，对于异常消费应当具备及时关停服务的能力。

2.2.7 大数据与人力资源管理

随着经济的全球化和一体化，国与国之间，企业与企业之间逐步演变成人才的竞争，谁拥有人才，谁就能够取得竞争优势。

对于企业，人力资源管理包括对于员工的全生命周期管理，包括员工招募、培训、考核、评价、晋升、降级、工资、福利、离职等。下面从识人、用人、育人三个方面进行分析。

1. 第一阶段：识人

识人，就是发现适用于企业的人才。识人可以分为从企业外部发现并招募人才和从企

业内部发现并选拔人才两种类型。对于从企业外部招募人才，传统的做法是中华英才网、猎聘网等人才中介、朋友圈推荐的方式。对于从企业内部选拔人才，通常是采用竞聘、领导推荐等方式获得。

大数据时代，企业可以采用社交网站作为获取人才信息的新渠道，通过建立企业人才库，从不同来源收集人才信息并建立人才全视图，构建人才评价模型，对人才进行多维度打分，作为人才评价的参考和依据。

2. 第二阶段：用人

用人的目标是实现人尽其才，发挥员工的特长和优势，实现企业与员工的双赢。当然，用人的同时也需要结合员工自身的主观发展意向。

大数据记录了员工的业绩信息，可以作为任用的依据。同时，由于一个人的职业生涯可能就职于多家企业，这时应当尽可能收集员工全部从业经历，以便发现员工的技能和特长，将其与工作需求结合起来。

3. 第三阶段：育人

在当今时代，科学技术大大改变了社会生产与生活方式，社会变化比以往更快，社会的专业化分工越来越细，需要企业员工不断地学习新知识、掌握新技能才能跟得上时代。

为了使得员工获得企业需要的知识和技能，企业可以制订培训计划，让员工快速成长，提高企业的整体竞争力。

培训讲师队伍对于学习型企业来说非常重要。企业可以借助大数据构建培训师知识库，培训知识库的数据源可以是以往培训师信息，包括培训课程、培训对象、培训效果等，也可以从企业内部培养业余培训师。

2.2.8 大数据与财务管理

人才对于企业非常重要，但是与人“才”相对的另一个“财”也非常重要。人才为企业创新发展提供智力支持，而“财”则为企业发展提供资金支持，两者对于企业都是非常重要的。从服务型企业的一把手通常主要负责“人”和“财”也可以看出两者的重要性。

对于社会来说，资金是社会资源配置的工具和手段，对于企业来说则是企业内部各种资源配置的工具和手段。从资金的运动方向看，资金包括收入和支出两个方向。企业为了

为社会提供产品或者服务，首先需要消耗企业内部的各种资源，从财务的角度看，各种资源的消耗产生各种成本支出，然后企业再通过让渡产品或者提供服务，从客户那里获得收入和利润。

资金如水，可以渗透到企业生产经营的各个环节。从成本流看，企业融资需要消耗资金成本，需要向供应商支付材料费和服务费，为渠道商支付的销售佣金，向员工支付工资和奖金等。从收入流看，企业通过销售产品或者提供服务可以获得收入，因为对外投资可以获得资金利息，因此设备或者场地出租获得租金收入等。可见，无论是企业的利润收入还是企业成本支出，都是存在多种方式的。

企业的财务活动包括两个层次。在操作层次，财务专员需要按照财务会计准则进行应收和应付的管理；在管理层次，企业财务经理会按照要求形成各种财务报表，为企业生产经营决策或者外部监管机构所使用。

企业生产经营的目的是获取利润，因此成本效益是企业财务分析的主要考量点。成本效益可以通过多个维度来完成，比如产品维度、项目维度等。比如，企业需要预测或者评价某产品的成本效益情况，需要计算该产品的成本和该产品的预期或者已产生的收入。如果企业采用了项目的管理模式，那么就需要以该项目为中心，计算项目的成本及其收入。

传统财务分析的方法是关注财务发生的结果，通常采用资产负债表、利润表（损益表）、现金流量表来分析企业的财务运行情况。同时，由于企业所有权与经营权的分离，要求企业将财务数据向企业所有者公布。企业财务分析的第二个用途是指导企业的生产经营决策，比如某产品或者某项目是否具有成本效益，是否值得做？如果值得做，其预期收益是多少？

会计科目是对会计要素的具体内容分类核算的科目，可以记录企业的收入和支出情况，是财务管理的基础，可以形成资产负债表、利润表、现金流量表等。以会计科目为基础的财务会计记录了企业发生的收支数据，再结合企业项目、产品、人员等信息，可以从多个维度来观察企业的财务情况。

财务分析结果好比企业生产经营结果的一个“快照”，只能看到“结果”，不能看到“过程”，虽然能够满足企业外部监管机构或者投资者的需求，但是难以满足生产经营管理人员的科学决策需求。为此，业界提出了基于活动的成本分析方法，俗称 ABC（Activity Based Cost）成本分析法。通过会计科目与业务活动的影射，可以实现财务与业务的双向透视，如图 2-2-12 所示。

在企业的生产经营过程中，借助财务会计来记录企业的各种收入和支出情况，包括产

品销售、服务提供、场地出租等收入数据以及原材料采购、工资奖金支付、佣金支付、利息等支出数据。借助信息系统，以上数据可以从项目、产品、部门等多个维度查看，分析判断企业财务风险。

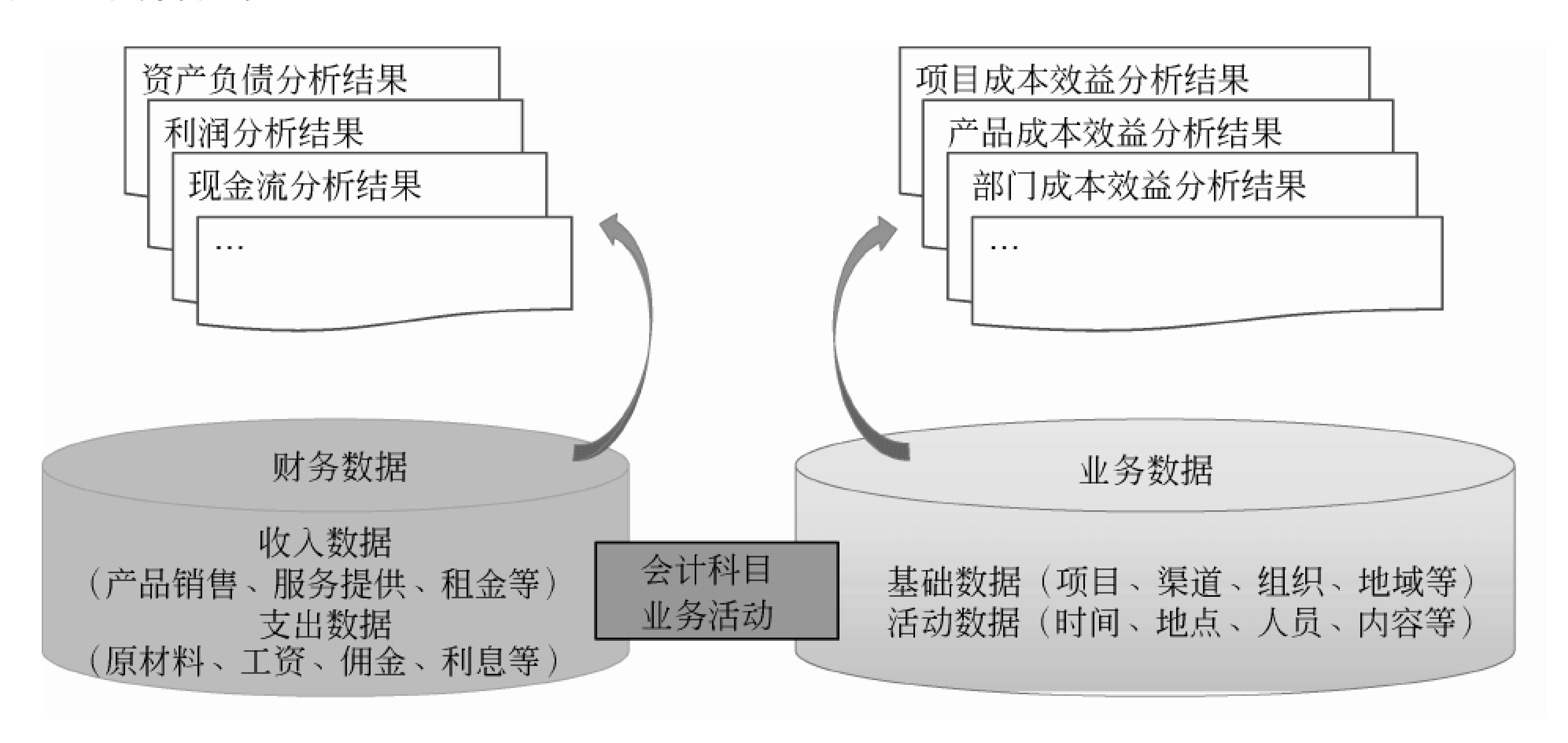


图 2-2-12　基于财务与业务的一体化分析

通过基于会计科目的财务核算，可以查看企业的财务表现，但仅仅可以看到某个时段内的“结果”，无法看到“过程”。如果将业务与财务贯通，就可以实现财务与业务的双向透视，帮助企业管理者发现企业业务活动中消耗的成本（人工工资成本、资金成本、资产折旧成本等），进行更加科学的资源配置。

2.2.9　大数据与资产管理

资产是从价值角度对资源的定义，资产范围包括有形的厂房、机器设备、工具、材料等，也包括无形的知识产权、商标、品牌等。随着移动互联网的发展，数据逐渐成为企业重要的无形资产。

对于有形的资产，可以借助大数据进行价值评估，提高资产管理的准确性。由于机器设备等有形资产会在使用过程中产生损耗，因此通常会通过折旧法来计算资产的现值，折旧方法包括直线法、工作量法、加速折旧法等，但是这种方法都具有天然的不足，企业可以借助机器设备使用过程中的运行记录、维修记录、保养记录等来辅助判断资产的价值，

做到对资产价值更加准确的评估。

2.2.10 本节内容小结

制定正确的企业发展战略的前提是"知己、知彼","知己"是企业应当能够及时、准确地掌握企业内部人、财、物等各种资源情况和能力情况。"知彼"是要掌握企业所处的外部宏观环境情况和竞争对手发展情况。大数据可以帮助企业汇聚来自企业内部和外部的各种数据，通过对比分析，发现问题和差距，进而制定符合企业发展的中长期发展战略。

企业发展战略明确了企业资源配置的目标、重点与步骤，成为企业基础设施建设的指南针，同时，企业基础设施建设又是产品构成和服务提供的物质基础。大数据可以帮助企业更好地完成规划设计、招标采购以及运行维护。在规划设计阶段，可以基于用户业务使用大数据来判断应用的价值和用户的价值；在招标采购阶段，可以基于不同来源的产品数据和用户产品使用评价数据，辅助完成采购产品的质量评价、配置对比、价格对比，利用大数据选取性价比高并且适合于企业的产品和服务。在运行维护阶段，可以根据基础设施的运行情况来制定基础设施优化和退出计划，并为基础设施的招标采购提供数据支持。

产品是贯通企业前后台的核心元素。前台反映了产品的市场特征，包括产品面向的客户群体、营销渠道、产品价格、服务方式等；后台反映了产品的资源特征，包括产品的构成、成本、生命周期等。通过分析产品的市场实施推广情况，辅助产品的定义。通过分析产品的价格、销量、单位成本、成本对象等，完成产品的成本效益分析，确定产品的市场进入、渗透、推广或者退出策略。

在社会商品极其丰富的买方市场，客户关系成为企业生存和发展的关键因素。从企业与客户关系的生命周期看，客户关系包括建立、维系、挽留、终止、恢复几个阶段，企业应当利用大数据，及时发现客户关系所处的状态并采取相应的措施。

供应商为企业生产经营提供输入，其提供的产品和服务决定了企业为客户提供的产品和服务的质量和速度。渠道商则帮助企业销售产品和服务，其产品的销售情况决定了企业应当制定什么样的产品开发策略，为客户提供什么样的服务。在价值网络时代，企业应当具备敏捷地响应外部环境变化的能力。大数据可以帮助企业及时、准确地掌握供应品市场情况和产品销售情况，以便企业做出正确的产品生产决策和供应品采购决策。

企业为客户提供产品和服务的同时，需要向客户收取费用。对于服务型企业，主要是

在客户的业务使用过程中进行计费和收费。对于客户规模大，业务复杂的服务型企业，需要具备费用实时或者准实时计费的能力和风险控制能力。企业可以借助大数据，提高业务使用记录的采集能力、计费能力、费用实时查询能力以及实时的风险控制能力，以便提升客户服务体验，降低企业收入流失。

科学技术是第一生产力，而掌握先进科学技术的是人才，未来社会的竞争必然是人才的竞争。因此，如何发现、吸引、选拔、留住人才成为企业提升竞争力的关键。企业可以利用大数据，设计人才选拔模型，寻找和构建适合企业发展的人才队伍。

企业的目标是获取利润，因此财务的“财”与人才同样重要。财务管理基础的会计科目，企业通常以一定的会计周期为单位统计形成财务报表，典型的财务报表包括资产负债表、损益表和现金流量表。财务报表是静态的，无法反映企业生产经营过程中的收支情况，因此需要通过业务与财务的映射，实现基于活动的成本管理。企业可以利用大数据，将财务数据与业务数据关联起来，从多个维度、多个环节透视企业的财务情况。

对企业而言，资源反映了物的使用属性，而资产则反映了物的价值属性。从财务的角度看，企业资产经过使用，随着时间的流逝会发生折旧，为了科学地衡量企业的资产状况，尤其是对于那些专业的资产管理公司，会通过资产的维修记录、保养记录、运行记录等数据来评估资产的价值，以便降低企业整体经营风险。

总之，大数据可以应用于企业战略、运营管理等多个方面，企业可以基于业务活动的不同阶段，想象大数据可以为企业能力带来的提升。

2.3 对号入座：定位大数据发力点

立足于业务过程框架和业务过程块，不仅能够有利于快速发现新的大数据服务，又便于从业务角度来管理越来越多的大数据服务。

前面已经从时间维度和空间维度，对企业业务过程进行了框架设计，将企业的各种业务活动进行分层分类管理，形成了既相互独立又相互联系的业务过程集合体。

企业以发展战略为指引，完成了从建设到运营的业务活动的实施。大数据的出现，为企业更好地完成业务活动提供了燃料，有了大数据服务支撑的业务活动，将会更加快速、科学地支持企业决策，提升企业的整体竞争能力。

下面就以业务过程框架为出发点，分析企业业务活动是如何利用大数据来提升自身能力的。下图 2-3-1 是企业业务过程一级框架，标有数字编号的矩形框是大数据服务可以支持企业业务过程的区域。

基于业务过程框架的大数据应用如图 2-3-1 所示。



图 2-3-1　基于业务过程框架的大数据服务覆盖区域

下面就以企业业务过程框架为指引，分析企业不同的业务过程域可能需要的大数据服务。

2.3.1　市场营销和提供管理

企业实施市场营销的最佳时机是企业有机会接触客户的时候，因此企业应当抓住与客户接触的大好机会，利用大数据，提升营销、销售以及服务能力。

客户通过打电话、上网、去实体营业厅等多种方式来获取企业提供的产品和服务，客

户与企业之间的交互行为包括网页浏览、故障申告、业务投诉、咨询建议、业务使用、费用支付等。企业应当充分利用这些接触点，把握时机，利用大数据进行针对性营销、销售和服务。

企业可以结合客户所在的位置、使用的访问设备、接入的渠道等场景，推荐适合客户的产品或者服务，实现企业产品或者服务与客户需求之间的精准匹配。

市场营销与提供管理过程需要的大数据服务包括：

（1）浏览时进行实时个性化的推介。例如，搜索引擎记录了客户近期的搜索历史，当客户浏览网页时，可以投放与客户搜索关键词相关的广告。

（2）结账时实时的个性化产品推介。客户为产品或者服务支付费用时，企业可以基于预先对客户偏好的分析结果向客户推荐产品或者服务。

（3）在线互动时实时的个性化产品推介。例如，当客户向企业咨询问题时，企业可以为客户推荐可能喜欢的产品或者服务，实现交叉营销或者向上营销。

（4）基于位置的实时的个性化产品推介。例如，当通信用户离开归属地并在拜访地停留接近一个月之前，电信运营商可以为用户推荐新的套餐。

（5）基于使用的实时的个性化产品推介。当客户正在使用企业提供的业务 A 时，可以为其推荐其他业务。比如当客户使用企业提供的主机托管产品时，可以为其推荐云计算产品和云安全产品。

（6）基于设备的实时的个性化产品推介。企业可以根据用户访问的应用，结合用户移动终端当前的能力，发现当前终端设备能力的不足，并为用户推荐具备支持该应用能力的终端。用户只需单击链接就可以进入产品展示界面，客户可以查看推荐的终端信息并一键下单。

（7）基于浏览历史的智能化广告。客户浏览历史包括浏览网页、停留时长等信息，企业可以根据浏览历史预测客户的消费倾向，为其推荐符合其偏好的商品。

2.3.2 服务开发与管理

服务是企业对资源的封装和编排后形成的能力。企业为了实现敏捷地响应外部市场需求，需要首先将资源封装为服务，然后通过对服务的重新编排，快速形成新的业务。服务可以分为不可再分的原子服务和由多个原子服务组合而成的组合服务。

例如，电信运营商的宽带接入业务是由三个服务编排后形成的，这三个服务分别为：

基于线路的线路配置服务、基于端口资源的端口配置服务以及基于账号资源的用户认证服务。从资源到服务的实现过程如图 2-3-2 所示。

大数据作为企业的新型无形资源，同样可以像企业有形资源那样，通过封装和编排，满足企业内部需要和对外开放需要。笔者将服务开发与管理过程的大数据服务分为电信运营商大数据服务和虚拟运营商大数据服务两大类。

企业对外提供的大数据服务主要是要考虑隐私与法律问题，可以通过匿名、统计数据开放、数据审批等方法来解决隐私触犯以及由此带来的法律纠纷问题。

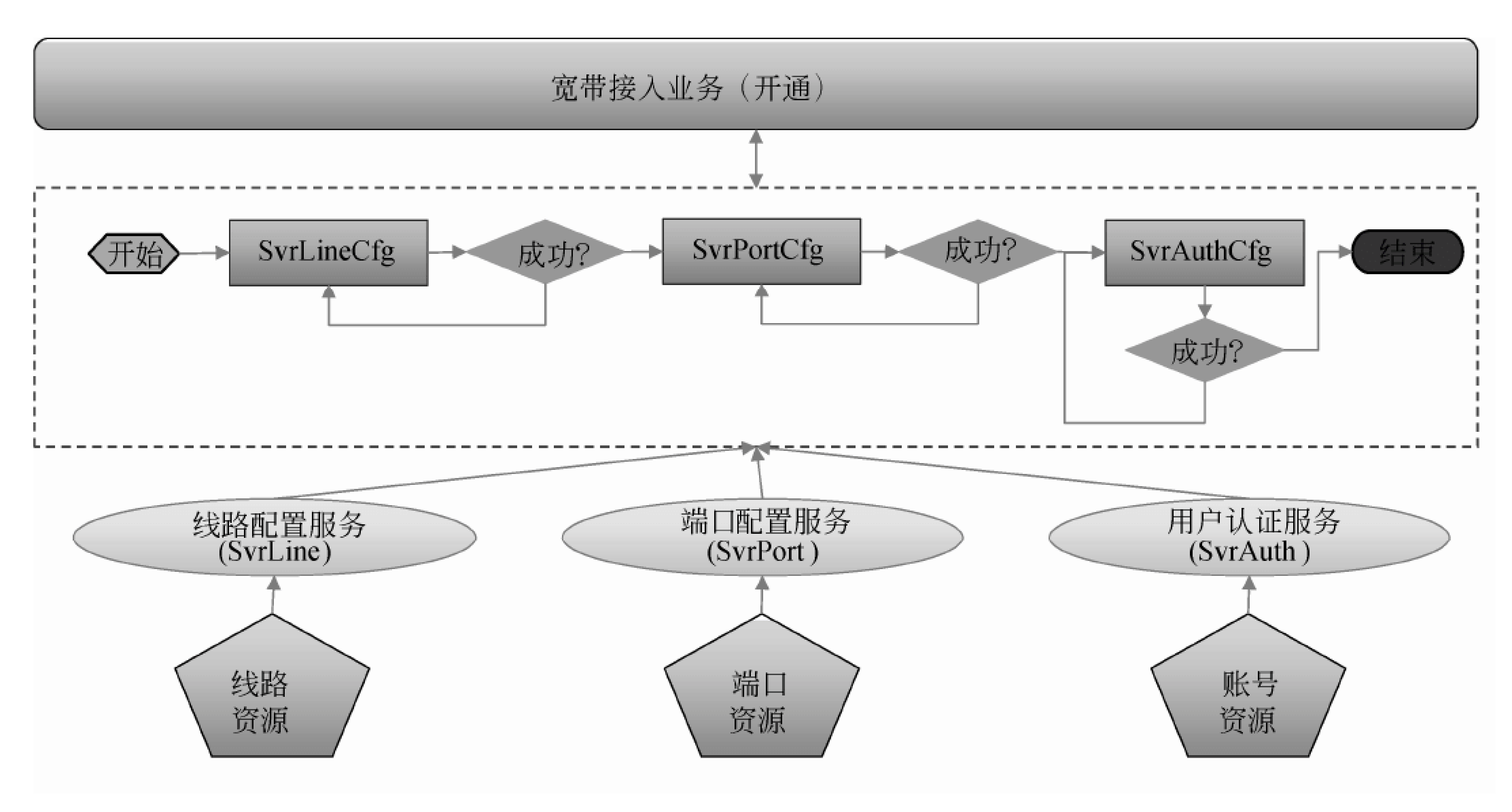


图 2-3-2　宽带接入业务从资源到服务形成的过程

2.3.3　资源开发与管理

资源是企业提供产品和服务的核心支撑，而企业的资源总是有限的。例如，通信网络资源的开发需要大量的资金投入，可以基于客户价值实现资源的最优配置。

对于一个发起新订单的客户，如果他是新客户，则可以将与他（她）的订单相似的客户对网络的影响作为参考点进行网络容量预测。如果他是一位老客户，就可以通过分析该客户以往订单对于资源的影响来调整资源容量，以便为客户提供可靠的服务，提升客户感知。

基于策略的能力管理主要针对那些资源配置和变更难度大的企业，大数据可以分析客户当前或预期行为、客户价值等，然后资源规划者可以针对不同用户的使用行为设计不同的策略，软件定义网络（Software Defined Network,SDN）就是一个例子。

以电信运营商的网络资源为例，其包含的大数据服务有：

（1）基于价值的网络规划。可以基于用户业务的使用行为形成的大数据，分析网络影响的用户特征，用户价值水平，根据用户价值来确定网络建设或者扩容的优先级。

（2）新订单影响分析。新订单势必对现有资源容量、配置等造成冲击，因此应当利用网络资源数据，对新订单的影响进行分析。

（3）基于策略的容量管理。通常情况下通过增加硬件和软件资源来提升系统的容量，借助网络质量、客户价值、客户使用偏好、客户支付偏好等大数据，预测客户新的需求对于系统容量的影响，就可以实现基于策略的容量管理，实现资源的优化配置。

2.3.4 服务实施

服务实施过程属于售中阶段，其作用是完成客户业务的开通。服务实施一般包括订单处理和安装实施两个环节，企业可以利用订单处理、客户自助安装、企业上门安装中的经验，提高订单转化成功率，提高安装实施的客户满意度。服务实施过程的大数据服务包括：

（1）增加在订购过程中的转化。在业务开通的过程中，客户可能会因为开通时间长、竞争对手提供更优惠的产品资费等原因而放弃开通，这时企业可以利用客户所在位置、使用设备等现场数据来分析和预判客户取消订单的概率并预先采取措施。

（2）减少订单处理过程中的错误。通过预先分析和测试订单处理过程中涉及的环节，减少订单处理过程中的错误。

（3）客户自助安装优化。通过对历史的客户订单数据，包括安装位置、安装人员、成功与否等进行分析，预测客户自助安装胜任度、优化设备配送流程、主动提醒客户、测试客户是否成功安装。

（4）现场技术人员优化配置。通过对于历史安装或者维修数据的分析，分析不同技术人员擅长的技能，将任务分配给最合适的人员，实现现场技术人员的排班管理。

（5）现场技术人员到场时间优化。根据以往技术人员解决任务的时长、日程安排等数据预测现场技术人员到达客户现场可能的时间或者延迟时间，对到场时间进行优化，并及时通知客户，提升客户感知水平。

2.3.5 客户关系管理

当客户使用企业的产品后，根据客户对产品或者服务的使用情况，企业应当对客户进行关怀、维系、挽留等。

企业提供人工服务的成本比提供自助服务的成本高很多，为了降低企业客户服务成本，企业应当尽可能引导客户采用自助渠道解决问题。

企业通过分析客户资费、网络信号质量以及从社交媒体等收集到的数据，提前预测客户可能咨询的问题，并主动推送给客户，以减少客户使用人工服务的次数和时间，降低企业运营成本。

电信运营商可以为客服代表提供网络分析数据，以提升客户服务能力。网络分析数据包括多个维度，比如客户所在区域、行动路线等空间维度，还可以是语音通话、视频通话、网络浏览等媒体和应用维度。企业根据客户所处时间、地点、行动轨迹、所使用的终端、应用等对网络性能进行分析，以便客服代表更好地为客户服务。

客户关系管理过程需要的大数据服务包括：

（1）个性化的实时互动。通过大数据分析，可以对客户可能向企业获取的服务进行预测，以将客户问题迅速引导到正确的流程和人员，快速为客户解决问题并提高企业运营效率。

（2）增强客户自助服务的有效性。自助服务可以大大降低企业运营成本，但是自助服务毕竟是机器提供的，智能化程度比人工服务要差，企业可以借助大数据服务，分析企业在提供自助服务过程中存在的问题并进行优化完善，通过提高自助服务水平降低整体运营成本。

（3）主动关怀。通过对下单、支付、业务使用等记录的分析，主动发现客户在不同环节存在的问题并实施主动的客户关怀，比如通过分析电信网络质量记录，对客户使用中的不便进行短信致歉。

（4）在恰当的渠道和时间实施主动关怀。通过大数据分析，发现客户接受服务的渠道和时间并主动实施客户关怀，比如分析发现客户倾向于通过短信渠道与企业互动，那么企业可以优先采用短信渠道与客户互动。

（5）基于糟糕关怀体验的主动关怀。通过大数据分析，发现那些具有糟糕关怀体验的客户并实施主动关怀。

（6）用于客户维系的流失风险预测。通过大数据分析，发现具有流失倾向的客户并进行主动维系，防止客户流失。

（7）用于客户维系的流失动机预测。通过大数据分析，发现客户流失的动机，为维系客户提供参考和依据。

（8）用于客户维系的个性化推介。通过大数据分析，可以发现客户流失的动机以及可能流失的客户群体，对比实施相应客户关怀后而没有流失的客户，找出防止客户流失的个性化方案。

（9）在网络故障期间或之后实施主动关怀。信息服务提供商通过大数据分析，发现网络出现故障，则对于网络影响范围内的客户实施主动关怀。

（10）采用网络体验分析提升客户关怀能力。这是信息通信服务提供商特有的大数据能力。通过大数据分析，可以建立一个全面的、不同视角的客户网络体验画像，包括不同位置、不同时间段、不同使用类别等视角。

2.3.6　资源管理与运营

运行中的资源会发生故障并影响到客户对于企业提供服务的使用，企业可以利用大数据分析以往类似故障产生的原因和解决方案，并应用到新的网络故障检测和修复过程中。

网络带宽资源总是有限的，尤其是无线网络资源。当用户数量多而带宽资源有限时，通常会发生网络拥塞，就像公路上塞车一样，车多道路资源少，车自然走不起来。在这种情况下，企业可以利用大数据来获取客户价值、流失风险系数等数据，根据这些数据来为用户分配网络带宽，保证高价值用户具有更好的网络使用体验。

电信运营商的网络设备发生故障后用户无法使用通信服务的一段时间内，电信运营商可以利用大数据来分析因网络故障受到影响的客户以及客户所处的位置、使用的应用、客户的价值、流失风险等，根据这些数据来进行网络修复并通知受影响的客户群。

资源管理与运营过程需要的大数据服务包括：

（1）网络故障定位和恢复。企业可以根据网络告警、网络性能、网络运行日志等网络大数据，实现网络故障定位和恢复，提高网络的自动修复能力。

（2）基于价值的实时拥塞管理。通过对网络进行 DPI 操作、客户 ARPU 等大数据分析，可以发现客户正在使用的业务类型和客户的价值，然后可以根据客户的价值来提供不同质量的网络服务。

（3）客户实时降级管理。通过大数据分析，可以实时发现网络资源情况和客户价值，将低价值客户自动转移到低价值网络中，这样高价值客户就可以享受更好的网络服务，降低了企业的总体运营成本。

2.3.7 合作伙伴关系管理

作为企业合作伙伴的分销商、批发商、零售商等在产品销售中发挥着非常大的作用，企业可以利用大数据来调整激励手段和佣金规则。

合作伙伴关系管理过程需要的大数据服务为合作伙伴价值优化。通过大数据分析，可以调整激励计划、佣金规则、结算规则等，通过有效管理企业与合作伙伴的关系，提高企业销售能力。

2.3.8 计费与收入保障

企业的收入有多个来源并且经过多个处理环节，如果不能够准确处理会对企业造成收入流失，对于计费收费中出现的错误也应当及时调整并通知客户。

企业可以利用大数据来分析与收入相关的数据源及数据处理过程，发现存在的无主记录和错误记录，避免收入流失。

计费与收费保障过程所需的大数据服务为收入保障。通过大数据分析，对客户在业务使用、采集、计费等所有环节进行监控、分析和预警，及时发现问题和解决问题，防止收入流失。

2.3.9 企业战略规划

制定企业战略规划决定了企业发展的方向和道路，而市场战略是企业战略的排头兵。企业可以利用内部和外部数据来制定市场发展战略。

市场观察对象和内容包括竞争对手产品、新产品、技术发展趋势、来自社交媒体的评论等。

企业战略规划过程所需的大数据服务有市场观察等。通过大数据分析，可以掌握企业竞争对手、其他企业市场空间、产品市场接受度等。

2.3.10 企业效率管理

业务过程执行效率的高低关乎企业的运营成本和客户感知，因此优化业务过程，提高业务过程执行的成功率非常重要。

企业可以利用大数据来发现业务过程执行失败的原因并及时解决问题，也可以通过自动适应的方式来提高业务过程执行的成功率。

企业效率管理过程所需的大数据服务包括业务过程优化等。通过大数据分析，可以发现企业业务执行过程中存在的问题，实现自动化的业务过程改进。通过业务过程优化，可以提高企业运营效率，提升企业内部员工和外部客户的感知水平。

2.3.11 财务和资产管理

据估计，企业大约有 3%的欺诈事件发生，欺诈事件为企业带来了很大的经济损失。

企业可以利用大数据从欺诈案例中掌握欺诈模式并通过暂停业务、提醒等方式来减少因欺诈带来的损失。

财务与资产管理过程所需的大数据服务包括防欺诈管理等。大数据分析可以从以往欺诈案例中提取出欺诈模式，可以对欺诈进行预测，并通过阻止、提醒等方式实现反欺诈。借助大数据分析，可以使得企业对于欺诈行为的预测更加准确。

2.3.12 本节内容小结

本节以企业业务过程框架为指引，分析业务过程所需的大数据服务，是一种从业务视角出发寻找大数据服务的正向思维。这种方法可以更加明显地看到大数据服务对企业业务活动的支持产生的价值和作用。

本节仅仅列举了几个典型的大数据服务，实际上企业根据数据源的多寡可以形成很多创新型的大数据服务。大数据服务可以帮助企业更好地制定发展战略，支持企业更好地制定建设和运营决策。应用需求主要来自于外部市场需求和内部管理要求，而大数据服务需求更多地依赖于企业管理者的经验和想象力。

2.4 能力落地：大数据服务数据源及其关键实现活动

数据源是大数据服务的“根”，决定了大数据服务的能力，可以基于可能获取到的数据源，初步确定实现大数据服务的关键活动。

前面首先基于企业业务过程框架，对大数据应用进行了畅想，然后再以企业业务过程框架为支点，分析了其所需的大数据服务。下面从数据的视角，分析这些大数据服务所需的数据源及大数据服务实现的关键活动。

2.4.1 聚集大数据：发挥资源聚合效应

不同组织具有不同的职能，因此每个组织的信息系统功能及其产生的数据也势必不同。大数据的特征之一就是数据的多样性，而要发挥大数据服务的作用，首先要集成不同来源的数据，才能发挥整体能力。

例如，对于社会中存在的某个自然人来说，其行为轨迹通常会发生在购物、人际交往、购物、旅行、就餐等活动中，而这些轨迹则分别被商场、超市、旅游公司、餐饮店、银行等机构所记录。自然人的生活轨迹如图 2-3-3 所示。

如果能够尽可能多地获取到关于某个自然人的数据，就能够更加准确地把握他（她）的行为取向。对于销售产品的企业来说，就能够更好地推荐产品，提供服务。不同来源的数据汇入大数据资源池如图 2-3-4 所示。

站在企业的角度，数据越完整越好，这样才能更好地反映个人或者组织的全貌。可以说，企业采集的数据越全面，数据的活跃度越高，企业就拥有了更多的“资产”。当然，如何不对这些“资产”进行挖掘，那么即使这些数据是“金子”，也不会发光的。为了展示大数据服务形成过程，首先需要掌握大数据服务的数据源及其关键实现活动。

2.4.2 行业通用数据源及关键实现活动

大数据服务形成的基础是来自不同渠道、不同信息系统的数据源，通过对数据源的分

析可以更加清晰地看到大数据服务对于数据的需求。

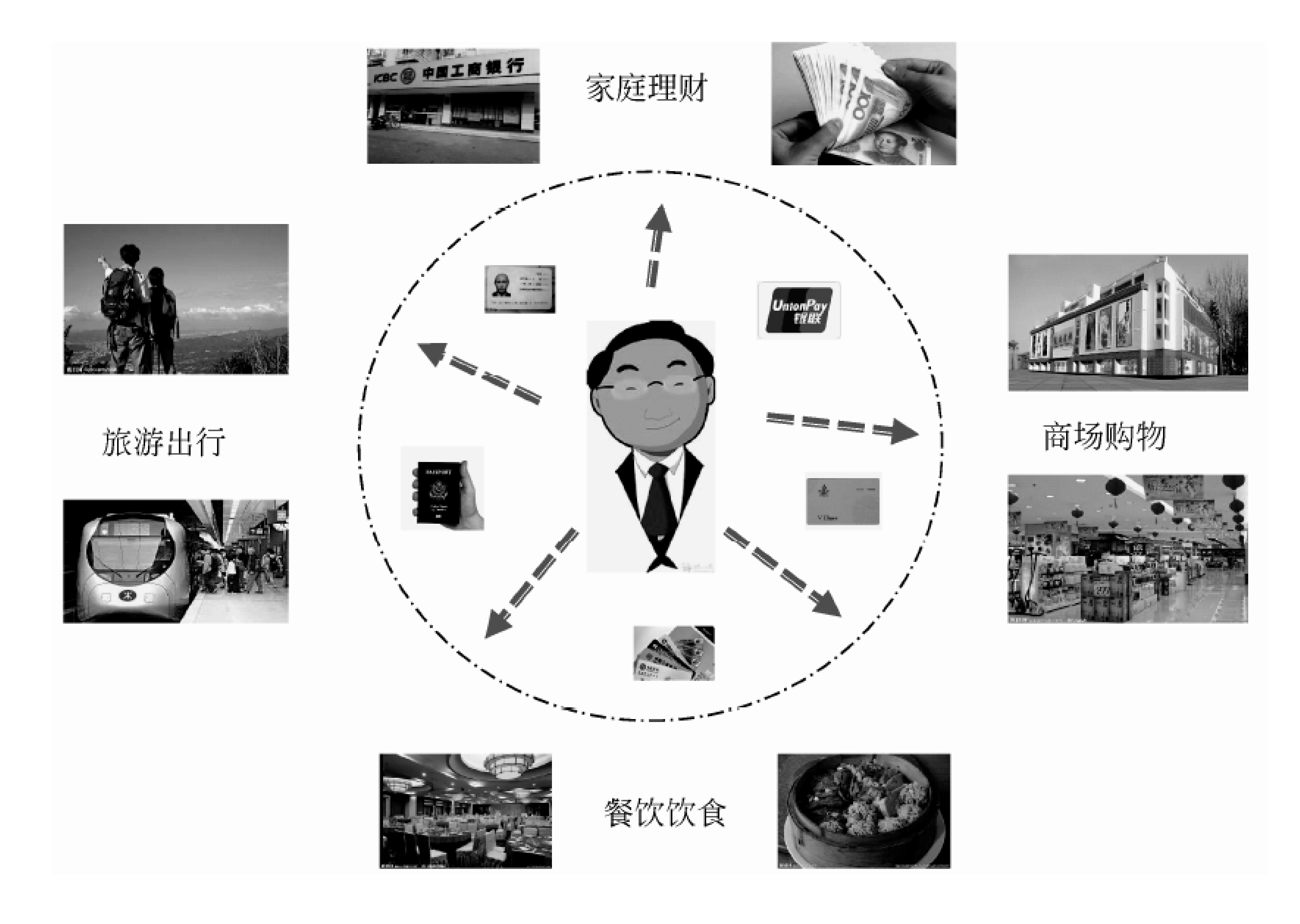


图 2-3-3　自然人的生活轨迹

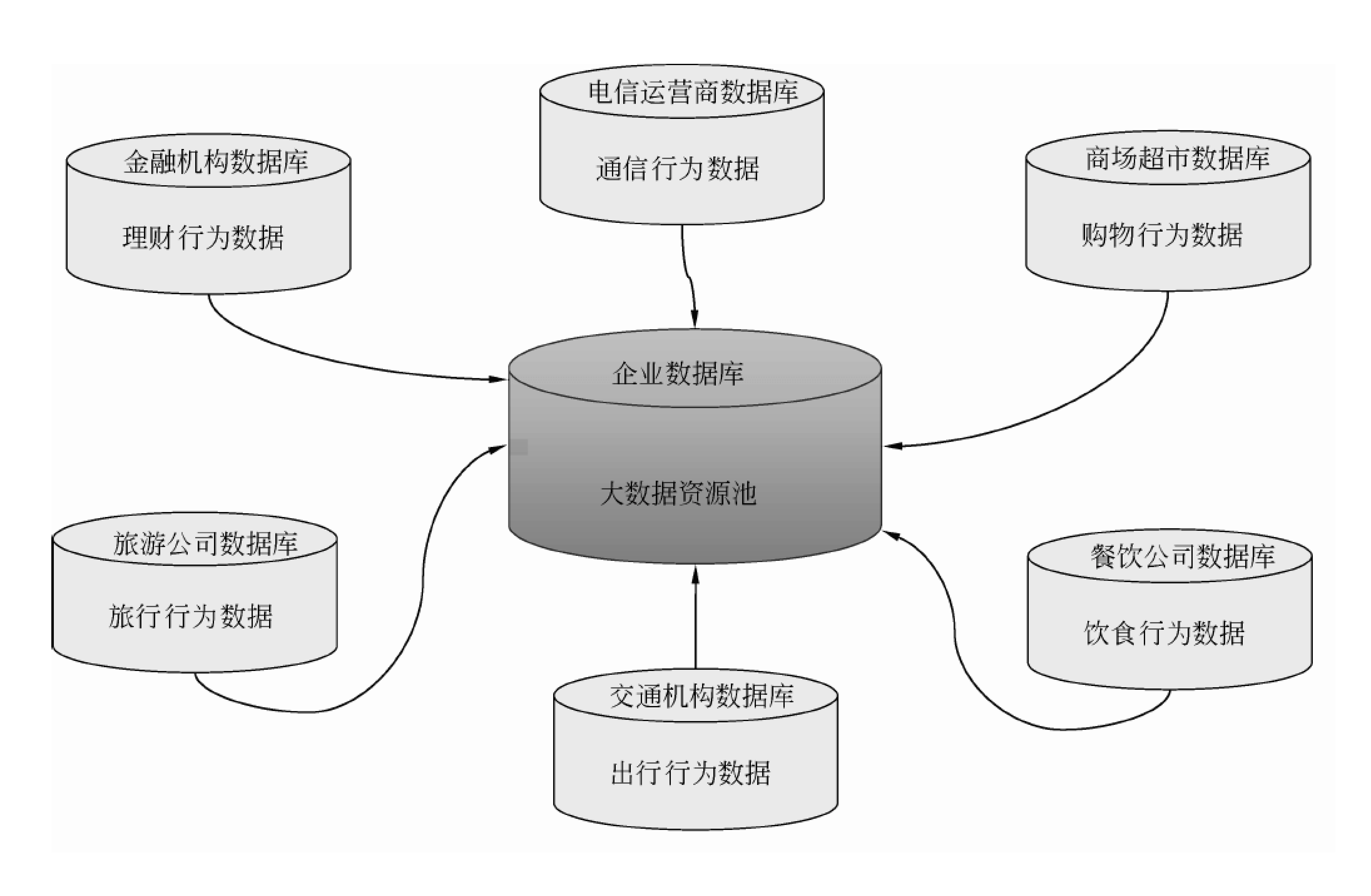


图 2-3-4　不同来源的数据汇入大数据资源池

数据源分为必选和可选两种类型。必选数据是大数据服务形成的前提条件，可选数据可以让大数据服务具备更强的能力。本节以电信运营商大数据服务所需数据源为例。

1. 浏览时进行实时个性化的推介

1）数据源需求

（1）必选数据源。

产品目录、可用性与适用性规则、当前正在或者刚刚浏览的商品。

（2）可选数据源。

CRM 数据、使用和计费信息、商品购买历史、业务使用数据、客户位置信息。

2）关键实现活动

S1：收集客户购买、使用、支付、位置等客户相关信息；

S2：收集产品、产品目录等商品相关信息；

S3：分析客户偏好，基于可用性与适用性规则，实现客户需求与企业商品供给的有效匹配，将最适合客户的商品展示在客户浏览区域附近。

2. 结算时实时的个性化产品推介

1）数据源需求

（1）必选数据源。

产品目录、可用性与适用性规则、当前正在或者刚刚浏览的商品。

（2）可选数据源。

CRM 数据、使用和计费信息、商品购买历史。

2）关键实现活动

S4：收集客户信息，获取 ARPU 等反映客户购买力的数据，为上行销售或者交叉销售做准备；

S5：收集企业产品信息，并根据可用性和适用性规则进行替代商品或者补充商品推荐。

3. 在线互动时实时的个性化产品推介

1）数据源需求

（1）必选数据源。

产品目录、可用性与适用性规则、当前交互场景。

（2）可选数据源。

CRM 数据、使用和计费信息、商品购买历史、网上浏览历史、业务使用数据、接入终端使用数据。

2）实现活动

S1：获取客户购买、业务使用、接入终端使用、支付、浏览等行为数据，计算客户的消费能力、购物偏好等。在线交互的渠道包括 Web、电话、自助终端、POS 等；

S2：当客户选购好商品，开始结算时，根据客户的消费能力、购物偏好等推荐其他可选商品或者更高级的商品。

4. 基于位置的实时的个性化产品推介

移动互联网时代的到来，企业可以掌握客户的位置与移动轨迹，因此可以基于客户位置数据来进展针对性营销。

1）数据源需求

（1）必选数据源。

产品目录、可用性与适用性规则、移动位置信息、客户列表。

（2）可选数据源。

通话详单、社交媒体记录、网上浏览历史。

2）关键实现活动

S1：收集产品和产品目录数据；

S2：收集客户近期通话、社交网络以及网络浏览的历史数据，分析客户近期的位置变化规律；

S3：根据客户所在位置进行产品或者服务推荐。比如对通话行为分析后发现客户在三个月内经常有不在归属地的漫游通话，在社交网络中也有晒漫游地的照片等，就可以判断这个客户近期经常出差。航空公司可以为其推荐航班，电信公司可以为其推荐适合出差的套餐。

5. 基于使用的实时的个性化产品推介

企业可以结合客户使用的终端或者应用推荐产品或者服务。比如客户上网流量超过某个值后，电信运营商可以立即为客户发送短信推荐更高的产品套餐。

1）数据源需求

（1）必选数据源。

产品目录、可用性与实用性规则、业务使用数据、终端使用数据。

（2）可选数据源。

CRM 数据、使用和计费信息、购买历史。

2）关键实现活动

S1：分析客户的行为和使用规律；

S2：将每个客户当作一个个体看待；

S3：利用反馈结果持续地提高营销准确性。

6. 基于设备的实时的个性化产品推介

1）数据源需求

（1）必选数据源。

CRM 数据、购买历史、产品目录、网络与服务库存量数据、产品性能数据。

（2）可选数据源。

客户流失动机预测、订单数据。

2）关键实现活动

S1：分析客户当前使用的设备，包括设备型号、设备网络制式、设备能力等。

S2：分析客户每天如何使用设备，比如何时开关机、何时上网、何时打电话等。

7. 基于网页浏览历史的智能化广告

1）数据源需求

（1）必选数据源。

网页浏览历史，包括每页停留时间以及带有时间戳的执行动作。

（2）可选数据源。

CRM 数据。

2）关键实现活动

S1：分析客户的网页浏览行为，包括浏览的网址、停留的时长、单击的链接等。

S2：基于客户历史网页浏览行为推介相关的产品。

8. 移动用户行为货币化

1）数据源需求

（1）必选数据源。

通话记录单、位置与移动信息（来自移动网络或 GPS）。

（2）可选数据源。

设备特征（例如手机品牌、操作系统）、应用类型以及使用情况（浏览历史）、用户身份数据、消费记录（例如预付费最高消费或者后付费每月费用）、社交媒体及其他开放数据。

2）关键实现活动

S1：分析移动用户的行为特征；

S2：将移动用户行为应用于市场营销活动或者与第三方企业的合作事项中。

9. 产品定义与开发

1）数据源需求

（1）必选数据源。

订单数据、产品目录、CRM 数据、客户价值数据、使用和计费信息、产品性能数据、各渠道接触日志。

（2）可选数据源。

外部可以提升客户需求预测能力的数据源，例如竞争对手的产品；外部社交网络分析。

2）关键实现活动

S1：理解哪一款产品的绩效最好以及该款产品绩效好的原因；

S2：理解客户的偏好；

S3：深入理解竞争对手的产品。

10. 产品导入分析

1）数据源需求

必选数据源：客户位置信息，社交网络分析，语音呼叫分析。

2）关键实现活动

S1：分析引起产品成功和失败的因素；

S2：分析市场以及新产品在市场中的机会；

S3：通过历史数据预测新产品的绩效。

11. 产品性能优化

1）数据源需求

（1）必选数据源。

产品目录、可用性和资格规则、购物车内当前或者近期浏览过的商品、设备使用数据，（网络数据）、CRM 数据、使用和计费信息、购买历史。

（2）可选数据源。

客户位置信息、社交网络数据、语音呼叫分析。

2）关键实现活动

S1：分析当前或者以往产品成功和失败的因素；

S2：识别改进当前产品的机会；

S3：基于历史数据分析产品变更后的性能。

12. 产品购买倾向分析

1）数据源需求

必选数据源：为客户提供产品的历史记录，包括日期、时间、位置、渠道、成功与否等信息；产品购买历史，包括日期、时间、位置、渠道等信息；标识客户生命周期的事件，包括日期和时间，例如赔偿、付费、纠纷、投诉等。

2）关键实现活动

S1：分析每个客户与企业接触的偏好，包括接触时间、接触渠道等；

S2：预测客户购买产品的时间和地点。

13. 主动关怀

1）数据源需求

（1）必选数据源。

CRM 数据、使用和计费信息、支付历史、购买历史、网络质量。

（2）可选数据源。

用于增强对客户认识的社交媒体数据。

2）关键实现活动

S1：分析客户的历史交互行为背后的规律；

S2：预测对特定问题每个客户与企业互动的可能性；

S3：为客户关怀活动推荐或者执行适用的行动。

14. 主动关怀的最佳时间和渠道预测

1）数据源需求

必选数据源。数据和时间戳：通话日志、网页日志、外呼反馈、客户自服务接触、客户在线商店接触。

2）关键实现活动

S1：基于客户历史接触记录分析客户未来行为；

S2：将每个客户看作单独的个体；

S3：预测企业与客户最好的接触时间和渠道。

15. 基于糟糕关怀体验的主动关怀

1）数据源需求

（1）必选数据源。

通话日志、助理渠道的通话成绩单、非助理渠道的评价、社会渠道的客户投诉。

（2）可选数据源。

影响客户优先级的因素，包括客户 VIP 状态、客户生命周期价值、客户社会价值。

2）关键实现活动

S1：分析客户满意度低下的客户接触行为；

S2：对具有糟糕体验的客户群体实施主动关怀。

16. 基于未使用的主动关怀

1）数据源需求

（1）必选数据源。

设备使用数据、CRM 数据、理解客户活动的包检测。

（2）可选数据源。

客户投诉、客户流失动机预测、客户价值数据、网络性能数据、社交网络分析。

2）关键实现活动

S1：分析客户的业务使用行为和接触行为之间的联系；

S2：识别未使用企业提供的业务的客户；

S3：主动引导客户使用企业提供的业务。

17. 用于客户维系的流失风险预测

1）数据源需求

（1）必选数据源。

CRM 数据、使用和计费信息、购买历史、支付历史、网络质量、通话记录、来自协助渠道的通话记录。

（2）可选数据源。

来自社交媒体渠道的投诉。

2）关键实现活动

S1：分析客户流失的关键因素；

S2：预测客户流失的概率。

18. 用于客户维系的流失动机预测

1）数据源需求

（1）必选数据源。

CRM 数据、使用和计费信息、购买历史、支付历史、网络质量、通话记录、来自协助渠道的通话记录。

（2）可选数据源。

来自社交媒体渠道的投诉。

2）关键实现活动

S1：分析客户流失的动机；

S2：预测具有高流失概率的客户的流失动机。

19. 制订客户维系个性化方案

1）数据源需求

必选数据源：客户流失动机预测、历史维系记录、CRM 数据。

2）关键实现活动

S1：分析每个客户的行为；

S2：利用客户维系反馈结果不断优化改进客户维系方案。

20. 维系方案接受概率分析

1）数据源需求

必选数据源：历史供应品（日期、时间、位置、渠道、成功标识等）、购买历史（日期、时间、位置、渠道等）、标有日期和时间的客户生命周期事件（退款、支付、纠纷、投诉等）。

2）关键实现活动

S1：分析客户与企业接触的时间、渠道等偏好；

S2：预测客户接受维系方案的时间和地点。

21. 实时的客户降档管理

1）数据源需求

（1）必选数据源。

网络质量数据、CRM 数据、客户价值数据、使用和计费信息。

（2）可选数据源。

帮助更好地理解单个客户活动的数据（执行深度包检测、设备分析等）。

2）关键实现活动

S1：获取不同网络选项下的网络质量；

S2：分析客户行为模式；

S3：基于对客户预期行为的预判来预测哪种网络适用于客户。

22. 合作伙伴价值优化

1）数据源需求

必选数据源：合作伙伴管理数据、计费和使用事件、产品目录、网络数据、CRM 数据、购买历史。

2）关键实现活动

S1：分析当下激励计划、佣金规则以及结算方式的性能；

S2：预测激励计划、佣金规则以及结算方式的目标性能；

S3：设计改进的激励计划、佣金规则以及结算方式；

S4：在方案实施之前进行性能模拟。

23. 增加订购过程中的转化率

1）数据源需求

（1）必选数据源。

订单及相关元数据、订单过程中收集的客户数据、客户场景数据（位置、设备、物理基础设施等）、每个订单相关操作的事件数据。

（2）可选数据源。

客户居住地的社会经济信息。

2）关键实现活动

S1：分析客户的真正需求；

S2：改进订单流程，提升订单转化成功率。

24. 减少订单处理过程中的错误

1）数据源需求

（1）必选数据源。

订单及相关元数据、订单过程中收集的客户数据、客户场景数据（位置、设备、物理基础设施等）、每个订单相关操作的事件数据。

（2）可选数据源。

客户居住地的社会经济信息。

2）关键实现活动

S1：分析订单执行过程中出现错误的位置；

S2：基于执行过程中出现错误的订单影响到的客户的价值的高低进行优先级排序；

S3：主动或者被动地修复发现的问题。

25. 增强客户自助服务的有效性

1）数据源需求

（1）必选数据源。

CRM 数据、外呼渠道成功与否的反馈。

（2）可选数据源。

社交媒体渠道的投诉、订单数据。

2）关键实现活动

S1：分析自助服务渠道流程中的改进点；

S2：发现产品营销的主要改进点；

S3：利用双人测试方法发现自助服务渠道的界面改进点。

26. 客户自助安装优化

1）数据源需求

（1）必选数据源。

订单及相关元数据、订单过程中收集的客户数据、客户场景数据（位置、设备、物理基础设施等）、自助安装案例的呼入记录、外送设备的配送记录、来自客户设备的数据。

（2）可选数据源。

客户居住地的社会经济信息。

2）关键实现活动

S1：分析客户的自助安装技能水平；

S2：以最快的速度将安装材料配送到客户手中；

S3：确保客户已经成功完成安装。

27. 现场人员优化配置

1）数据源需求

（1）必选数据源。

现场技师目录、以往现场支持数据（技师必须做的工作、花费的时间、成功还是失败、什么产品错误）、客户特征及每个任务的现场场景。

2）关键实现活动

S1：分析每个技师的主要技术专长；

S2：确定能够解决现场特定问题的合适技师；

S3：通知技师需要为某项任务重点准备什么。

28. 现场人员到场时间优化

1）数据源需求

（1）必选数据源。

专业技师工作计划、每个工作的历史档案、现场技师工作起止的实时通知。

（2）可选数据源。

帮助技术评估的地图和交通数据。

2）关键实现活动

S1：持续获取现场每一位技师的任务进展状态；

S2：预测技师当天日程安排中其他任务的到场时间；

S3：当技师到达现场的时间变化时，主动通知客户；

S4：当技师无法按照预定时间到达客户现场时，需要为该技师重新安排任务。

29. 收入保障

1）数据源需求

必选数据源：计费和使用事件、网络数据、CRM 数据（如支付历史）。

2）关键实现活动

S1：检查未开票交易，即虽然客户已经使用业务，但是企业由于未知原因未收取费用；

S2：识别改进采集过程、减少投诉的机会；

S3：识别企业与合作、结算、漫游相关的可能得到改进的过程。

30. 个性化收费处置计划

1）数据源需求

（1）必选数据源。

可用的收费行为目录、以往收费行动、周期、要回债务的成功率、应用于特定客户的收费行动、客户特征。

（2）可选数据源。

客户居住地的社会经济信息，使用、购买、订购或者其他能够更好地理解客户行为的数据源。

2）关键实现活动

S1：将客户作为单独的个体看待，发现收回资金的最佳方式；

S2：对待那些可能无法联系的客户要快速采取行动。

31. 市场观察

1）数据源需求

必选数据源：企业战略计划、产品性能数据、企业市场营销机会、外部社交网络分析、外部技术数据、外部市场数据。

2）关键实现活动

S1：分析市场发展趋势；

S2：评估企业应当跟随哪一种市场发展趋势；

S3：基于外部市场环境和内部企业资源，设计企业战略发展规划。

32. 防欺诈管理

1）数据源需求

必选数据源：计费和使用数据、网络数据、位置数据、CRM 数据（比如支付历史）、销售点数据、供应数据、信用积分/历史数据、外部法律相关（调查、公安、司法等）。

2）关键实现活动

S1：从以往的欺诈案例中归纳出欺诈模式；

S2：基于事件预测可能的欺诈行为；

S3：采取阻止、提醒等方式应对欺诈行为。

33. 业务流程优化

1）数据源需求

（1）必选数据源。

CRM 和计费数据、订单管理和供应数据、市场营销与维系供应品数据、合作伙伴数据、客户投诉、服务问题、业务流程事件与数据、业务流程日志、社交媒体。

（2）可选数据源。

供应链、工人、天气、交通等。

2）关键实现活动

S1：发现失败的过程及过程改进的机会；

S2：预测可能会失败的过程并提前采取行动；

S3：在过程失败之前给出主动解决问题的建议；

S4：通过自适应过程改进，优化系统性能，实现企业成本效益的最大化。

2.4.3 电信行业数据源及关键实现活动

数据来源于不同的地方，不同行业也有着不同的数据特点，本节以电信运营商特有的数据源为例进行说明。

1. 促进预付费转后付费

1）数据源需求

（1）必选数据源。

CRM 数据、购买历史、产品目录、网络与服务库存量数据、产品性能数据、使用数据。

（2）可选数据源。

社交媒体数据。

2）关键实现活动

S1：分析客户从预付费转到后付费的原因；

S2：识别那些与预后转化具有同样特征的客户；

S3：定位预付费客户，并用正确的方案将其转化为后付费用户。

2. 网络故障期间或之后的主动关怀

1）数据源需求

（1）必选数据源。

客户投诉、网络故障数据、客户资料。

（2）可选数据源。

位置数据、网络性能数据、网络质量数据。

2）关键实现活动

S1：分析以往处理类似事情的成功因素；

S2：识别网络故障期间影响到的客户；

S3：推荐正确的行动方案，以免客户主动联系企业。

3. 基于网络体验分析的主动关怀

1）数据源需求

必选数据源：客户人口统计、网络数据（集中在 3～7 层网页日志）。

2）关键实现活动

S1：分析客户在特定的时间、位置、应用以及使用方式的网络体验；

S2：识别客户在网络使用体验中需要解决的问题；

S3：对客户在网络体验中出现的问题的严重性进行打分。

4. 网络故障定位与恢复

1）数据源需求

必选数据源：网络与服务库存量数据、网络故障数据、网元日志文件、网络性能数据、服务质量数据、客户投诉、现场测试、客户关怀代理记录、网络与服务使用数据。

2）关键实现活动

S1：基于历史网络故障数据，形成网络失败模式；

S2：基于历史网络故障数据，掌握网络故障恢复方法；

S3：基于网络故障模式分析结果，预测网络故障；

S4：推荐或者执行正确的网络故障恢复过程，解决网络故障。

5. 基于价值的实时拥塞管理

1）数据源需求

（1）必选数据源。

网络质量数据、CRM 数据、客户价值数据、使用和计费数据。

（2）可选数据源。

帮助更好地理解单个客户活动的数据（执行深度包检测、设备分析等）。

2）关键实现活动

S1：判断网络策略是否要求干预；

S2：识别客户当前的活动；

S3：基于业务活动，为每一个客户预测合适的调节水平；

S4：基于关键业务因子实施网络优先接入策略，例如客户价值，流失风险等。

6. 主动体验驱动的网络修复

1）数据源需求

（1）必选数据源。

网络数据（集中在 3～7 层）、CRM 数据、客户价值数据。

（2）可选数据源。

网页日志。

2）关键实现活动

S1：基于网络运行数据自动检测网络运行中存在的问题；

S2：自动实施网络问题修复；

S3：基于客户价值对问题解决顺序进行排序；

S4：主动将问题通知到客户。

7. 电信运营商数据货币化

1）数据源需求

必选数据源：所有可用的数据源，比如网元、BSS、OSS 等。

2）关键实现活动

S1：跨越企业多个管理域采集并聚合数据；

S2：构建数据使用方访问聚合数据的接口。

8. 虚拟运营商数据货币化

1）数据源需求

（1）必选数据源。

电信运营商内部与虚拟运营商客户相关的数据。

（2）可选数据源。

与虚拟运营商客户相关的所有可用的数据源，比如网元、BSS、OSS 等。

2）关键实现活动

S1：收集与虚拟运营商相关的数据；

S2：构建帮助虚拟运营商发现价值的大数据分析服务。

9. 基于价值的网络规划

1）数据源需求

（1）必选数据源。

网络质量数据、CRM 数据、客户价值数据、使用&计费信息。

（2）可选数据源。

用户提升对客户认识水平的社交媒体数据、用于增强对未来使用预测的外部数据源，比如新建的广场、公园等。

2）关键实现活动

S1：掌握客户在不同地点的价值分布情况；

S2：理解客户在不同位置的使用模式，比如在哪里上网，在哪里打电话等；

S3：预测客户在特定位置使用模式的变化。

10. 新订单影响分析

1）数据源需求

必选数据源：订单数据、网络质量数据、CRM 数据、客户价值数据、使用和计费信息。

2）关键实现活动

S1：分析用户历史行为，并将这些行为与其他客户行为进行关联；

S2：将新用户的预期使用行为分析作为新订单的一个部分；

S3：预测新用户对于网络质量的影响；

S4：根据预计的影响推荐网络变更方案。

11. 基于策略的能力管理

1）数据源需求

必选数据源：网络质量数据、CRM 数据、客户价值数据、使用和计费信息。

2）关键实现活动

S1：分析客户的行为；

S2：分析在当前策略下的网络性能；

S3：设计并模拟网络策略变更后的效果。

2.5 主要内容回顾

企业架构的目标是构建一个以企业发展战略为指导，有效连接业务与技术的、系统化的框架体系。通过这个系统化的框架体系，企业可以实现高效的运营，更好地发挥信息系统在企业发展中的价值和作用。

但是，仅仅依靠设计良好的企业架构还不能让企业具有很强的竞争力，在移动互联网时代，企业架构必须与大数据紧密结合起来，让大数据成为支持企业架构的不竭动力，为企业发展战略提供参考依据，为企业建设和运营提供决策支持，提升企业生产和决策过程的自动化和智能化水平，降低企业运营成本，提升企业整体运营效率。

企业的生产经营活动反映在不同的、相互联系的业务活动中。业务活动分为决策型和操作型两类，决策型业务活动则负责业务活动执行前的分析和判断，而操作型业务活动只负责执行，两者就像人的大脑和四肢，是“知”和“行”的关系。大数据服务的目标是支持企业决策型活动，企业的决策型业务活动就成为发现大数据服务的切入点，是一种从决策需求出发，正向发现大数据服务需求的方法。

为了实现对业务活动的有效管理，通常采用空间和时间两个维度相结合的方法，形成矩阵式的分层分类管理框架，即业务过程框架。相比于业务活动，业务过程框架能够更加清晰地定位和管理大数据服务。企业业务过程在时间维度和空间维度的交叉，形成了既相互独立又相互联系的过程块。以业务过程框架中的业务过程块为线索，可以快速定位大数据服务的发力点。

大数据服务在业务过程框架中能够实现对业务目标的支持，那么能够支持业务目标实现的关键是丰富、全面、高质量的数据源，可见，数据源是大数据服务存在的前提和基础，同时，要想实现大数据服务，同样需要清晰地把握大数据服务实现的关键活动。

总之，企业架构要想有力地支撑企业的战略、建设和运营，必须要借助大数据。大数据服务让企业决策更加全面、高效、正确、准确、自动化以及智能化，可以让企业快速感知外部环境变化，及时调整企业发展战略，提升企业运营效率，降低企业运营成本，提升客户感知水平。同时，大数据服务也必须与企业业务过程框架相结合，这样大数据服务才能有机地植入到企业架构中，更好地找到自身的发力点。可见，企业架构与大数据的“联姻”是企业在移动互联网时代的必然选择。

第 3 章
孕育：凡事预则立，不预则废

在“筑巢”阶段，企业以发展战略为指导，搭建了一套非常漂亮的房子，为爱情和事业打下了基础。有了坚固耐用可以挡风遮雨的房子，下一步就要恋爱结婚了，谁和谁恋爱呢，按照本书的故事安排，应当说是企业架构和大数据，经过一场浪漫缠绵的爱情长跑，企业架构和大数据终于走向“婚姻”的殿堂了。

企业架构和大数据“结婚”后，自然想到要有两个人的小宝宝，小宝宝还没有出生，小两口就开始从长计议了：将来把孩子培育成什么样的人才？如何设计孩子不同年龄段的教育计划，等等。

以上就是以个人成家立业的过程为喻，分析企业如何从“筑巢”、“联姻”一直到“孕育”的思路和方法。

正确的观念决定一切。企业架构与大数据联姻之后，首先要改变思维方式，树立正确的观念，从全局和长远设计大数据服务。

树立正确的观念重要，掌握正确的方法也非常重要，方法正确就会事半功倍，否则就会事倍功半。应当认识到大数据服务与面向操作的事务型应用具有各自的特点：前者对于数据的质量要求高，而后者对系统的响应性能的要求高，前者强调正确性，后者强调及时性和可靠性。

软件开发，架构先行。从整体上进行架构设计，可以让各个部分高效地协同配合，提高开发效率。按照从业务到技术逐步落地的思路，可以将大数据服务分为业务运营和应用治理两个层次。业务运营框架关注整体结构，而应用治理框架则关注落地过程。

数据模型是实现大数据服务的关键部分。为了看到数据从产生、加工到集成、汇总的全过程，将支持业务操作的数据模型和支持决策分析的多维数据模型连接起来，以便更加清晰地看到大数据服务这个“成功人士”背后的“贤内助”所做的默默无闻的支持和努力。

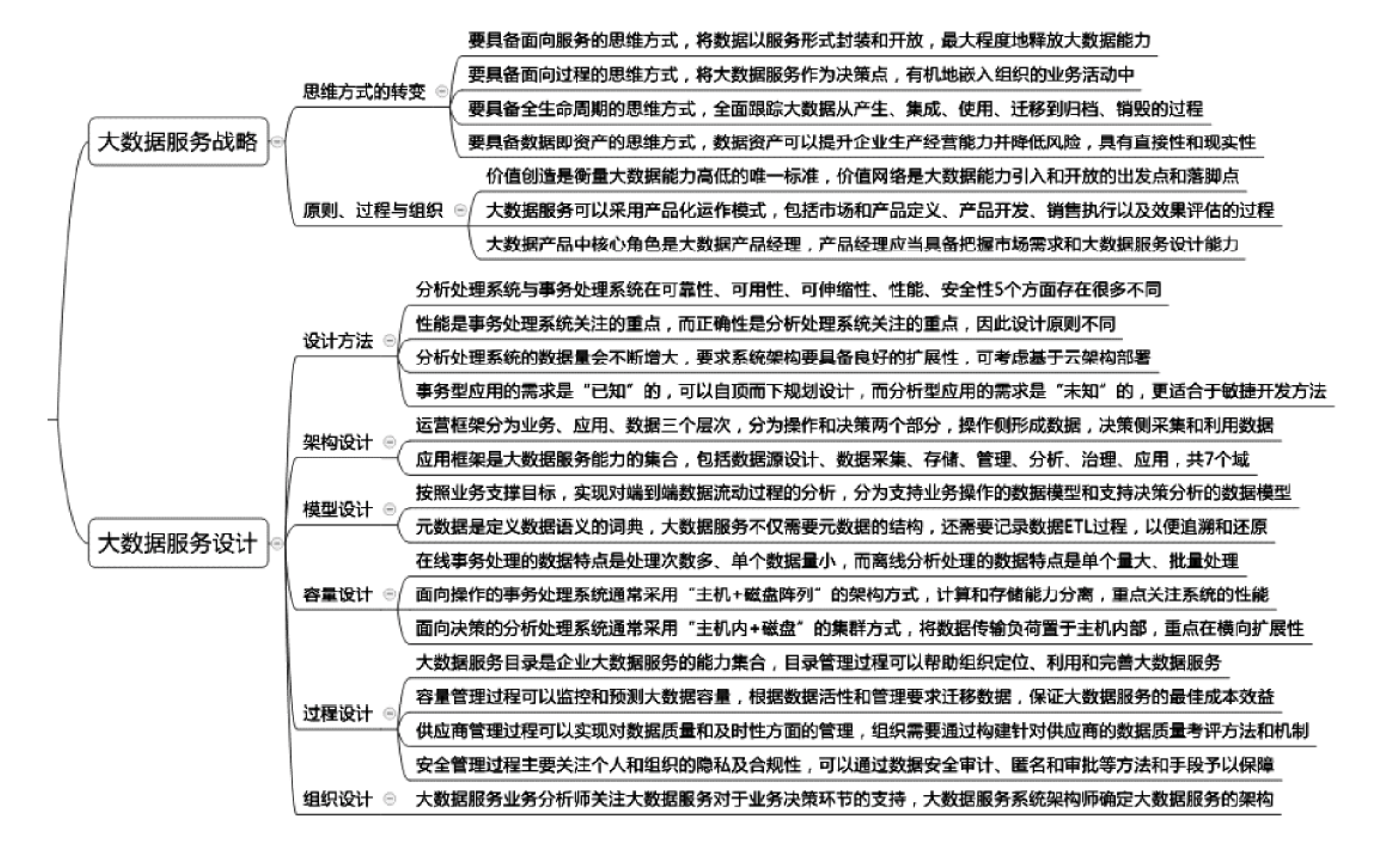
容量设计要满足大数据服务的正常运行要求，应当对不同类型的数据采取不同的容量管理策略，保证企业成本效益的最大化。大数据的特点之一是数据随着时间的推移不断增

长，因此应当满足计划期内大数据服务容量的需求，还需要能够监控数据的增长规律，及时扩容基础设施资源。

要根据数据的活性采取不同的数据管理策略，对于系统中一定时期内不使用的休眠数据，要及时迁移到低成本的存储区，节约大数据服务整体的存储成本，实现最佳成本效益。

大数据服务过程设计的目的是通过管理流程及时发现和解决问题。在大数据服务设计阶段，可以定义关注大数据服务业务需求和关注大数据服务架构实现的两个角色。

本章内容的思维导图如下：



3.1 大数据服务战略：大数据决定大未来

数据服务战略既是企业面向外部市场竞争的需要，又是企业释放自身能力的内在需求，是企业长远发展的必然选择。

人类社会的不同发展阶段其核心资产是不一样的。在生产工具落后的农耕时代，土地是社会的核心资产，战争大多因疆界而起。在工业时代，蒸汽机的出现大大提高了人类的生产能力，而石油、电力等新能源的发现和应用则推动了工业社会的高速发展，能源取代土地成为社会的新型资产，国与国之间的战争更多地体现为争夺“能源”的战争。自从人类发明了计算机和互联网，人类进入了信息社会，信息几乎渗透到生产生活的所有角落，而随着信息社会的不断发展，数据逐渐成为推动人类社会发展的新“能源”。

在当代，信息技术的作用从提高企业生产经营效率逐步转变为企业生产经营的主导力量，对于互联网企业，信息系统的建设水平则成为企业竞争力高低的重要标志。伴随着互联网的发展，人类活动的场所大部分迁徙到虚拟的网络中，通过手机、平板电脑、桌面电脑等接入设备，在网络上获取、分享、交换信息以及娱乐、购物等。

随着市场经济的发展，竞争引发了更加专业化的社会分工。通常将为社会提供商品和服务的组织称为企业，购买并使用商品和服务的个人或组织称为消费者，而将构建公平、有序的市场秩序的组织称为政府监管机构。

随着信息通信技术和互联网在全社会的深入应用，存在于不同组织的数据越来越多，越来越丰富：银行等金融机构存储了客户财产有关的数据，比如存贷款记录、消费记录、账户余额等；政府等公共事业机构则存储了民众身份、税务缴纳记录、交通违章记录等；电信运营商存储了客户通话记录、上网记录、短消息记录等；而提供各种服务的互联网公司则存储了客户搜索、浏览、评论等记录。当然，在用户的使用终端中，无论是桌面终端还是移动终端，同样会保存个人的使用行为记录。可见，只要是人类活动借助了信息技术，就会留下“痕迹”，成为信息社会永远的“记忆”。

那么，企业为什么要制定大数据服务战略呢？

首先，数据是社会生产生活的记录，蕴藏了个人和组织过往的“需要”，通常这些“需要”可以反映个人和组织的“偏好”，通常个人的这些“偏好”比较稳定，企业可以根据这

些“偏好”，看人下菜，预测需求，达到提高产品销售能力的目的。

其次，由于社会专业化的分工，数据分散在不同的企业之中，为了更好地利用数据，企业可以将内部数据进行封装，然后“走出去”，通过销售企业大数据资产来增加收入。同时，企业也需要把其他组织的数据“引进来”，通过聚合不同的数据，形成关于客户、供应商、合作伙伴、产品、渠道等的全景图，通过对这些管理对象的把握来提升竞争能力。

可见，大数据服务战略既是企业面向外部市场竞争的需要，又是企业释放自身能力的内在需求，是企业长远发展的必然选择。

3.1.1　大数据服务战略新思维

大数据时代，数据成为企业发展的核心资产。为了更好地发挥数据的价值和作用，企业需要将开发和实施大数据服务放到发展战略的高度，转变思维方式，立足长远和全局，将数据资产与企业生产和运营紧密结合起来。

大数据服务战略新思维主要包括面向服务的思维方式、面向过程的思维方式、全生命周期的思维方式以及数据即资产的思维方式。

面向服务是将大数据以服务形式进行管理，大数据服务本质上是数据资产的一种“能力”，这种“能力”既能够与企业自身的生产经营活动相结合，也可以嵌入其他企业的生产经营活动中。

面向过程指的是企业在进行顶层设计或者规划设计时，将大数据服务有机地嵌入企业业务活动之中，不要孤立地看待大数据服务，要通过大数据服务完成企业各个层次的决策任务。

全生命周期的思维方式要求企业从数据产生、集成、迁移、归档、销毁的全生命周期来全过程地观察，观察数据在生命周期每个阶段发挥的作用，主动将数据转换到新的状态，优化大数据基础设施资源配置，实现成本效益的最大化。

数据即资产的思维方式是将数据看作资产负债表中的新的资产项，原因是数据能够直接帮助企业提升生产和经营能力并降低企业风险。数据资产同机器设备、材料、办公用品等有形资产一样支持企业的生产与运营，数据资产相比知识产权、专利、品牌等无形资产更具有现实性。

1. 面向服务的思维方式

“服务”在生活中的应用非常广泛。员工为企业提供的劳动是一种服务，企业回报员

工的方式就是工资。企业为人们提供工资又称为支付服务，可见，“服务”是一个相对的概念，不能够孤立存在。

软件经历了面向过程、面向对象、面向组件、面向服务的发展过程，最终，软件如同人们日常生活中的水电一样，成为一种服务。下面看一下软件从面向过程、面向对象、面向组件、面向服务的发展历程。

面向过程阶段：为了解决软件危机，人们提出了不同的软件工程方法论。最初，计算机采用汇编等低级语言进行编程，由于这种编程方式效率低，为了解决这一问题，人们提出了面向过程的编程方法，程序以函数形式存在，函数定义中包含输入和输出参数，同时也有函数库的支持。面向过程阶段的典型编程语言有 C、Pascal、FORTRAN 等。

面向对象阶段：随着软件项目规模的日益扩大，支持项目的软件代码规模越来越庞大，面向过程的语言难以满足快速应对需求变化的要求。为了解决这一问题，人们提出了面向对象的编程方法。面向对象即采用面向现实世界对象的思维来进行需求的分析、设计、编码和测试。例如，现实世界中存在学生张三、李四、王五等，那么在软件设计时也把它抽象成一个学生类，这个抽象的类可以实例化为张三、李四这样的学生。学生的属性包含学生名称、所在年级、归属班级、所选课程等，学生的行为包括入学登记、选课、考试等行为。由于面向对象的方法将现实世界中的对象与软件中的对象对应起来，因此软件的可读性、灵活性方面都有了很大提升，软件不再是那么晦涩难懂的事情，与现实世界的距离更近了。

面向组件阶段：虽然面向对象的分析与设计方法解决了软件可读性和可维护性问题，但是软件的复用还是限制在程序代码一级，为此提出了面向组件的设计方法。组件好比一个机器零件，由于组件是一个独立的软件部件，可以嵌入外部环境中，通过接口的形式与外界交互，因此可以在机器代码层面实现软件的复用。参与组件规范制定的各方由于利益的不同，逐步形成了以微软公司主导的 COM 组件模型、Sun 公司主导的 EJB 组件模型以及对象管理联盟主导的 CORBA 组件模型。

面向服务阶段：组件虽然在很大程度上解决了软件的复用性问题，但是组件仍然是基于特定操作系统平台之上的复用，软件组件能力难以在异构平台之间调用，为此，人们提出了软件服务的概念，即以服务形式对软件进行封装并对外提供。软件以服务形式提供与消费，进一步提高了软件的复用性。

面向服务的架构（SOA），将 IT 资源以服务的形式进行封装，使得无论是企业内部还是组织之间，都能够通过 IT 服务进行顺畅的衔接。借助 SOA，组织可以像搭积木一样快

速地集成来自组织内外的各种服务，快速形成新的业务，对于企业，可以提升快速响应市场的能力。当前，信息系统的所有层面都可以以云服务的形式对外提供，包括基础设施即服务、平台即服务、软件即服务、安全即服务、桌面即服务、网络即服务等。

从面向过程到面向服务的发展路线如图 3-1-1 所示。

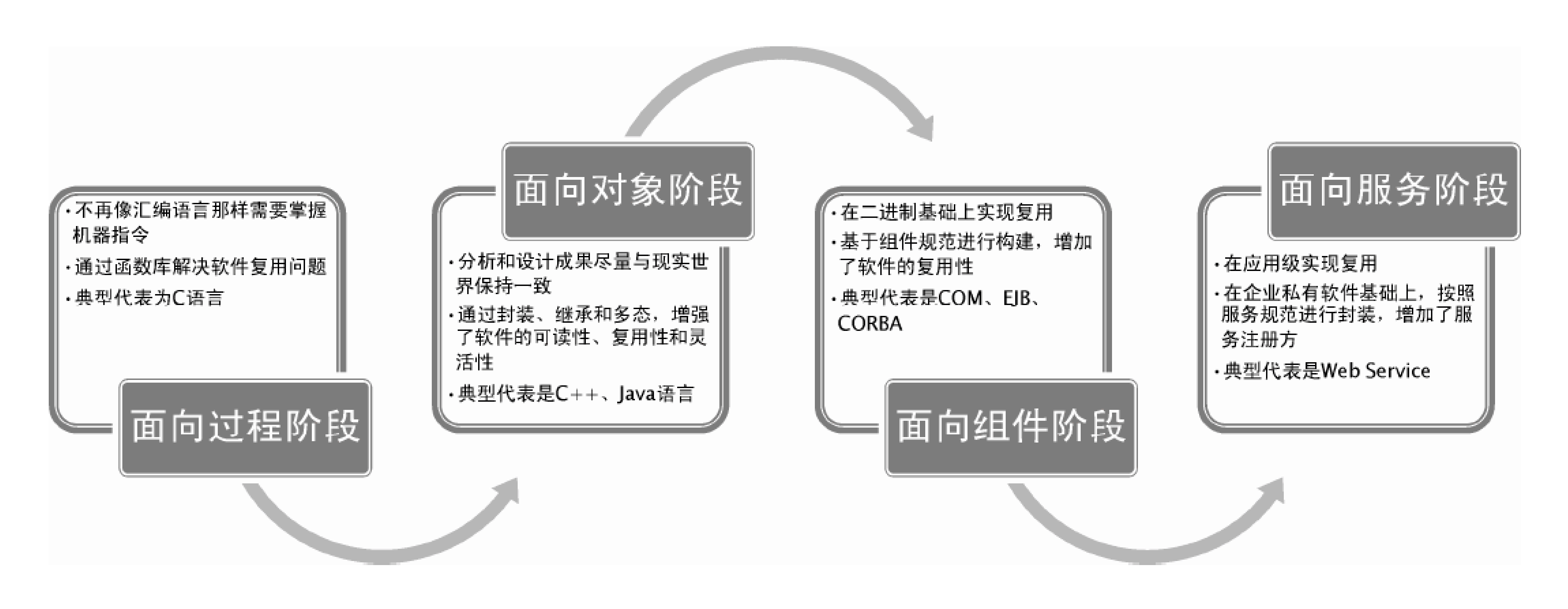


图 3-1-1　从面向过程到面向服务，软件复用性不断提升

在大数据时代，同样需要具备面向服务的思维方式。作为以提高组织决策能力的大数据服务，尽管并非所有大数据服务都像事务型应用那样需要快速的响应能力，大数据服务的响应时间也许是几个小时，甚至可以是几天或者一周，但是随着人们对大数据服务认识的逐步深入，许多大数据服务可以通过数据模型，逐步量化模型参数，使得大数据服务的决策结果逐步接近“准确”，因此，从长远看，企业仍然需要采用面向服务的思维方式，以各种大数据服务的形式将大数据的能力固定下来。

2. 面向过程的思维方式

企业如果采用职能型管理方式，即企业按照专业化分工分为市场营销、销售、服务、人力资源、财务等多个职能，会带来职能部门之间的沟通协同问题，职能部门往往会从自身利益出发屏蔽信息、争夺资源，这样的管理方式大大降低了企业的整体效率。

例如，企业要完成市场营销任务，按照职能型的管理方式，市场营销部门为了提升企业影响力，往往会在市场营销活动中做出超出企业能力的承诺，这些承诺会给工程建设部门带来很大的工期压力，为了赶进度企业通常会以降低产品和服务的质量为代价。

为了解决职能型管理模式存在的问题和不足，需要采用面向过程的管理方式。企业生

产、运营和管理的整体活动被分割为多个小的过程块，这些过程块之间协同配合来完成企业的某个任务。

以市场营销部和工程建设部为例，如果采用面向过程的方式，市场营销部在执行营销计划之前，过程规则要求市场营销部与工程建设部一起完成资源的核查，并为客户制定基于企业资源能力的市场营销计划，这样就不会存在两个部门因客户需求和资源供给不匹配而引起的问题。

大数据服务同样需要采用面向过程的思维方式。在大数据服务之前的商业智能，通常是开环的反馈机制，数据分析的结果与生产过程是分开的。采用大数据服务后，企业的生产过程与大数据服务采用闭环的反馈机制，无论这种反馈是由信息系统完成还是由人完成的。

虽然某些大数据服务只是呈现给企业决策者作为参考，并不嵌入系统的生产过程中，但是企业管理者的决策参考同样也是企业过程中的一个环节（或者称之为决策点）而已，只不过这个环节并没有由信息系统执行。图 3-1-2 是企业借助大数据服务实现决策的自动化的简单例子。

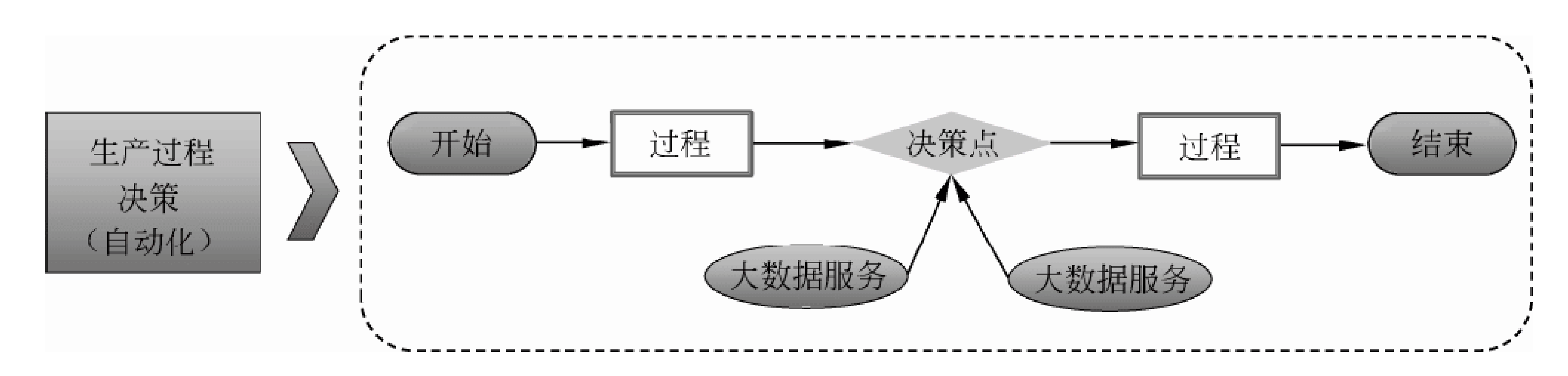


图 3-1-2　企业生产过程中的自动化决策方式

在企业生产经营过程中，为了提高生产或者经营效率，可以采用基于大数据服务的自动化决策方式，一个或者若干个大数据服务作为决策点的输入，根据输出结果和决策规则来确定决策结果。

然而，在许多情况下，有些决策比较复杂，还不能由信息系统取代，需要决策人员综合多种因素，借助头脑风暴等多种方式才能做出决策。在这种情况下，大数据服务可以作为企业决策者的决策输入。人工参与决策而不是大数据服务自动化做出决策，如图 3-1-3 所示。

在企业的生产经营过程中，人工参与决策的案例很多，尤其是那些战略层次的决策，这种类型的决策更多地依赖于决策者的直觉和经验，而影响决策的信息通常无法通过公开渠道获取。此外，目前大数据分析系统对于非结构化数据的处理能力有限，而某些决策更

多地依赖于文本、图片、语言、视频等多媒体数据，比如行业分析报告、政策新闻、谈话录音、购物录像等。

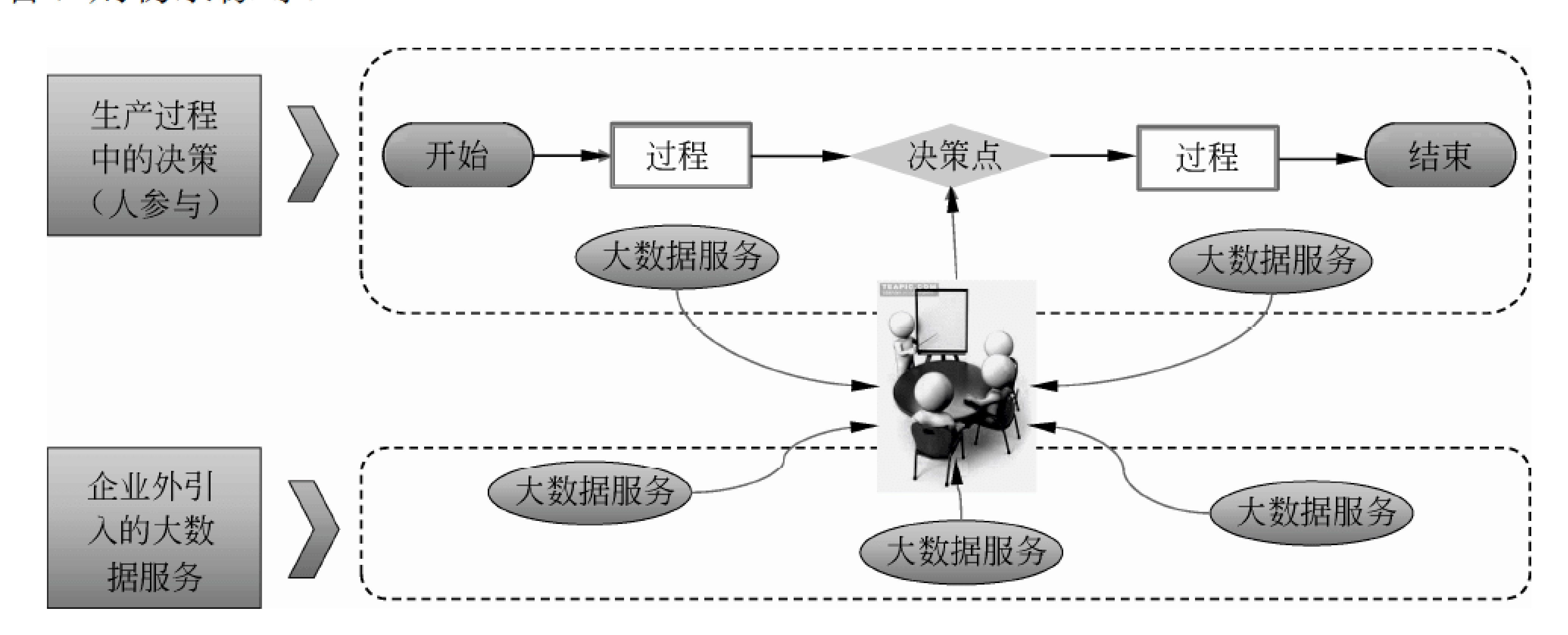


图 3-1-3　企业过程中的人工参与决策方式

3. 全生命周期的思维方式

宇宙中的任何事物，都具有从出生到死亡的过程，大到银河系的星球，小到地球上的微生物，无论其生命长短，都具有从出现到消亡的过程。

组织中的各种事物同样要经历从无到有，从有到无的发展过程。对于企业而言，其客户要经历考察期、形成期、稳定期和退化期四个阶段；企业的产品同样会经过投入期、成长期、成熟期和衰退期四个阶段；企业的设备资源会经过采购、使用、磨损、废弃等几个阶段；企业的员工会经过招募、入职、培训、提拔、解约等几个阶段；企业的信息系统会经过需求分析、架构设计、开发测试、上线运行、运维维护、下线退出等几个阶段。可见，企业的管理对象都要经历从引入到退出的全过程，将这一过程称为生命周期。

按照生命周期的思维方式进行分析，可以避免片面地、静止地、孤立地看待问题。反之，如果遵循全生命周期的思维方式，就可以全面地、运动地、联系地看待问题。

对于大数据，如果从数据的产生、存储、清洗、转换、丰富、使用、备份、销毁的全生命周期来进行分析，就可以更加完整地看到数据流动的过程，基于数据的价值实现数据的管理。

4. 数据即资产的思维方式

传统意义上的资产包括机器设备、材料、办公用品等，随着科学技术的发展，专利、

知识产权、企业品牌等知识资产成为社会的新型资产并逐步被人们接受。对于企业而言，资产成为企业成本效益核算的重要基础，企业可以通过资产负债表来反映生产与经营的风险水平。

随着信息技术在全社会的迅速应用，各种面向操作的事务型应用产生了越来越多的数据，包括来自大自然的环境数据、来自企业的生产经营数据、来自个人的生活数据以及来自政府部门的公共管理数据。当面向操作的事务型应用产生的数据规模很小的时候，人们通常只是对数据进行简单的查询和统计，目的也只是查找和定位以往的操作记录或者从总体上查看汇总数据。

随着数据规模的逐步增大，人们可以从大量的数据中发现许多规律性的东西，就好比人类通过日积月累，在生产生活中形成的经验和教训一样，生活阅历越丰富，就越能够体会到一些规律性的东西。在数据分析发展的过程中，出现了像 OLAP、数据挖掘、商业智能等研究方向，也可以证明人类在认识数据和利用数据方面不断地尝试和努力。

当前，随着传感器技术、移动通信技术和信息技术的不断发展，物联网和移动互联网逐步应用到公共管理、交通、医疗等行业之中。物联网和移动互联网的飞速发展，必将产生越来越多的数据，这使得大数据成为人类社会继煤炭、石油、电力之后名副其实的新型能源。同时，借助云计算，可以让大数据完全脱离物理资源的束缚，使得用户如同日常生活中获取水、电、气资源那样便捷地获取大数据服务。

数据成为企业新型资产主要还是取决于数据在企业生产经营中的价值和作用。当前，经济的全球化、一体化趋势日益增强，社会商品更加丰富，市场竞争更加激烈，产品同质化趋势越来越明显，企业为了生存和发展，必须寻找新的突破口和增长点，形成差异化竞争优势。由于数据来源于个人和组织的活动，更好地反映过去用户对于产品和服务的诉求，因此，企业可以以解决生产和经营中的问题为切入点，聚合社会上所有可用的数据，及时、准确地把握市场、产品、客户、员工、股东、监管机构等利益相关者的需求，提供个性化、差异化的产品和服务。

可见，在大数据时代，大数据的作用已经完全可以与机器设备、材料、办公用品等有形资产相媲美，与知识产权、专利、品牌等无形资产相比，大数据资产更具有现实性，企业可以运用大数据资产直接产生成本效益上的提升，通过生产经营能力上的提升来降低生产经营风险。因此，在物联网、移动互联网、云计算带动社会发展进步的时代，企业和个人必须认识到数据资产的重要作用，形成数据即资产的思维方式。

3.1.2 大数据服务战略原则

俗话说："没有规矩，不成方圆"，大数据服务同样需要指导原则。

与面向操作的事务型应用不同，大数据服务属于分析型应用，其特点更多地体现为探索性，因而大数据服务的风险也很大。如果能够从大数据中找到对企业决策有价值的信息，那么大数据服务的回报也是巨大的。如果说面向事务的应用的重点在于如何提高企业生产和经营的效率、降低成本以及提升客户感知，那么面向决策的分析型应用的关键则是减少失误。

面向操作的事务型应用是"正确地做事"，而面向决策的分析型应用是"做正确的事"，这就是两者在企业价值创造方面的不同之处。

1．价值创造原则

决定大数据服务是否有必要的前提条件是大数据服务能够为组织创造什么价值和多少价值，因此价值创造是大数据战略的首要原则。

对于大数据服务战略，其目的是提高企业利润，包括提高企业收入和节约成本支出两个方面。企业可以借助大数据服务来精确地掌握客户的需求并为客户提供满足其需求的产品和服务，通过提高企业的销售能力达到增加企业收入的目的。

一方面，企业也可以借助大数据服务，提升客户获取企业产品或者服务的便捷性和高效性，借助服务水平的提高达到提升客户感知的目的。客户对于企业感受好了，自然更愿意在企业购买更多的产品和服务。

另一方面，企业也可以借助大数据服务来发现企业业务流程中存在的问题，分析产生这些问题的原因并实施流程改进，通过流程的改进提升企业运营效率，这样可以间接地降低企业的成本支出。

大数据服务可以为企业创造价值，无论是直接创造还是间接创造。但是大数据服务设计时同样需要考虑大数据服务与企业已有业务过程的集成关系，保证大数据服务无缝地嵌入现有业务活动中。同时，不同的大数据服务在企业中的地位和作用也是不同的，应当分析大数据服务为企业带来的价值高低并进行优先级排序，优先保障那些对企业价值大并且具有紧迫要求的大数据服务。

2. 价值网络原则

任何企业都属于整个社会生产中的一个环节。企业以劳动、资金、原材料、知识等为输入，借助企业的采购、生产、运营以及管理活动，完成价值的创造，为社会提供产品或者服务，然后借助渠道将产品和服务传递给客户。这种从资源输入、价值创造到价值交付的过程，业界将其定义为“价值链”。由著名竞争战略专家迈克尔·波特提出的价值链模型如图 3-1-4 所示。

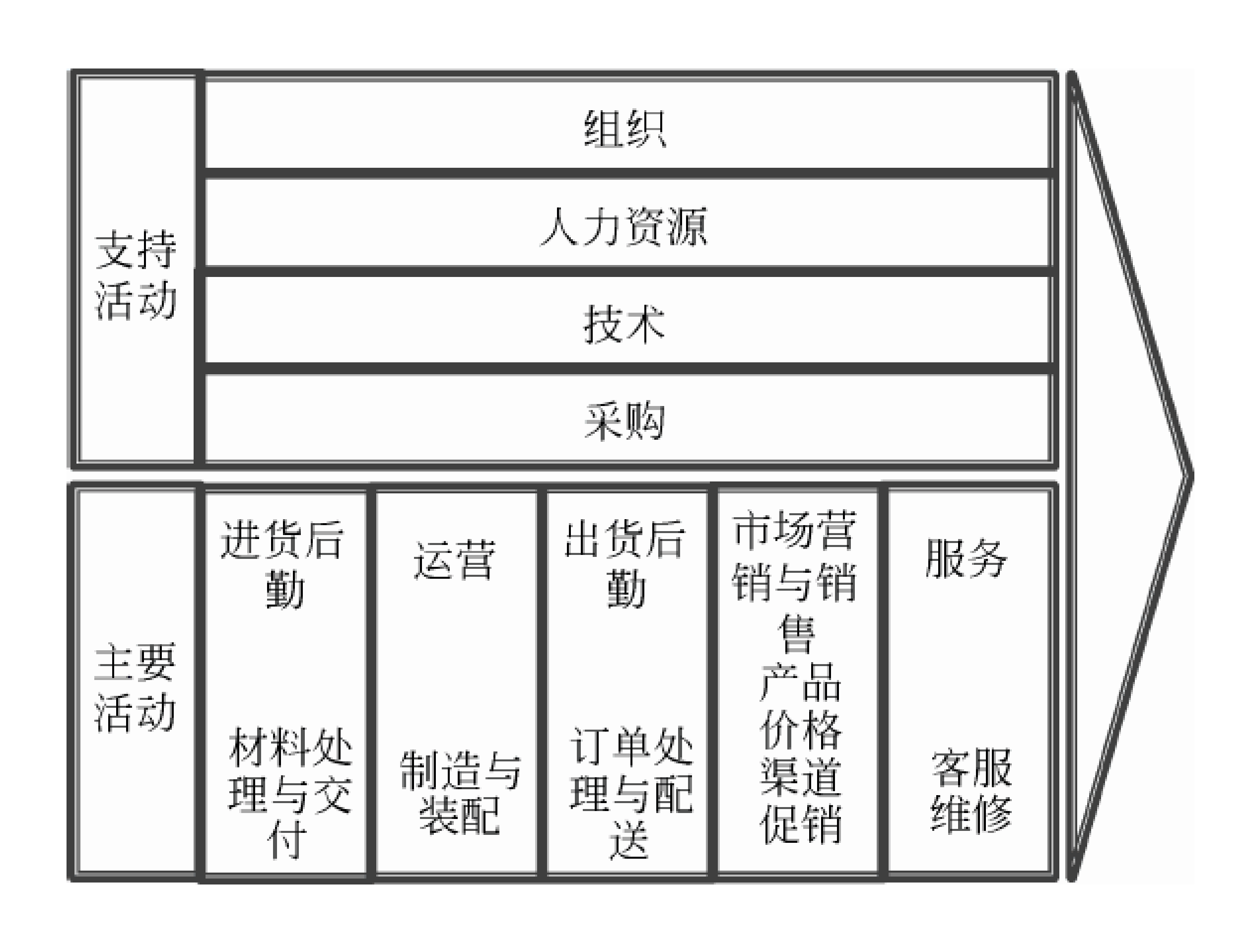


图 3-1-4 波特的价值链模型

从图 3-1-4 可以看出，价值链就像一根链条，形象地说明了企业输入、处理、输出的过程，随着近年来科学技术尤其是网络技术的迅猛发展，社会专业化分工越来越细，企业更需要与多个参与方协同，因此传统单一链条的价值链模式已经不再适用于现代企业，取而代之的是新型的价值网络模式。由客户、咨询师、设计师、服务提供商、领导者等角色参与的价值网络模型如图 3-1-5 所示。

价值网络时代更加强调企业之间的协同配合，企业应当取长补短，在竞合中找到自己在价值网络中的位置，不断提升企业自身的核心竞争力。

面对企业从价值链到价值网络模式的转变，企业在大数据服务能力的构建和服务能力的开放时，需要善于寻找和发现社会中已有的大数据服务，主动引入外部有价值的数据源，不断丰富和完善数据资产。同时，企业也应当积极将企业自身的数据资源以大数据服务的形式开放出去，最大限度地释放大数据服务的潜力，在竞争和合作中取得竞争优势。

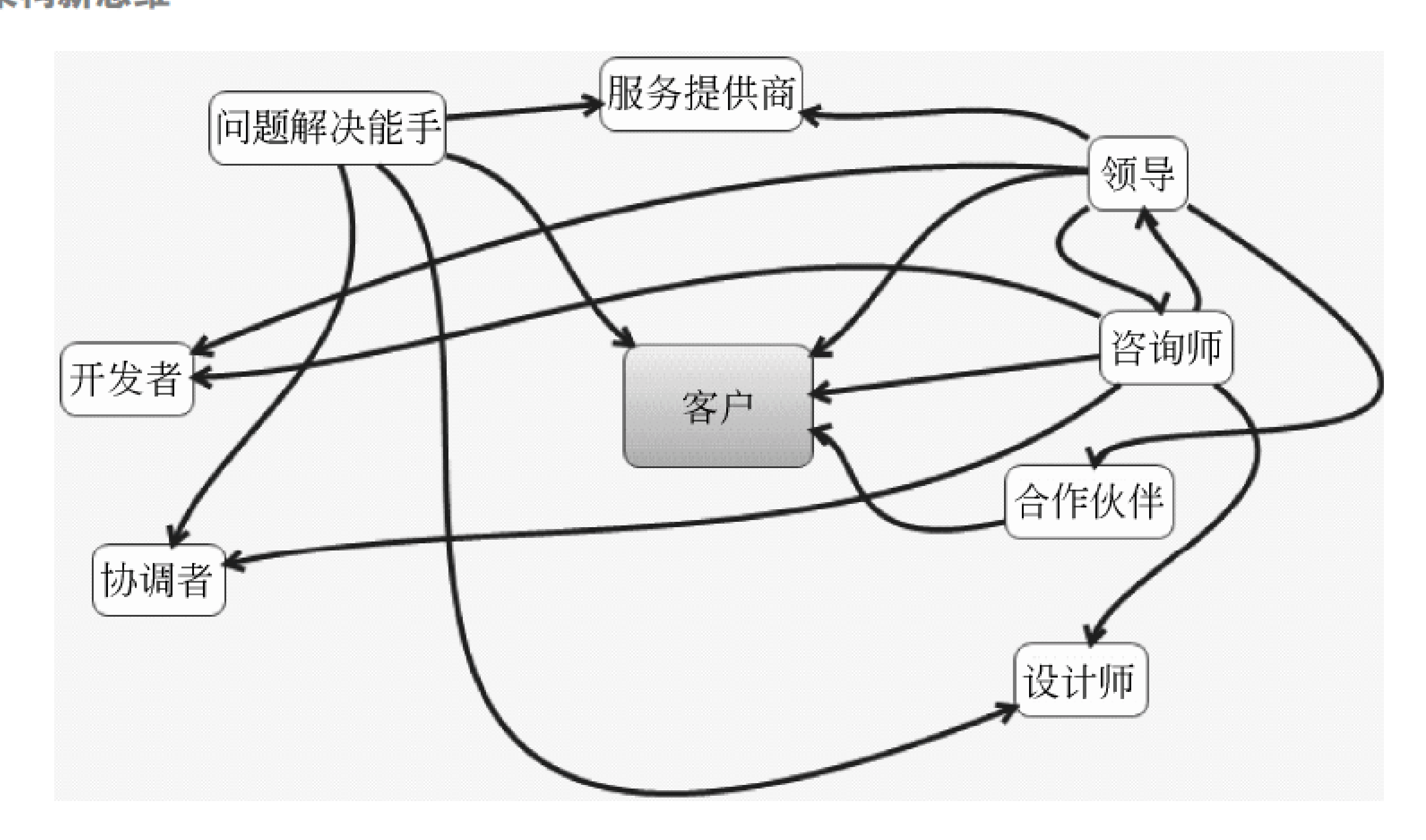


图 3-1-5　简单的价值网络模型

3.1.3　大数据服务战略过程

大数据服务战略包括市场定义、供应品定义、战略资产开发、战略执行以及战略评估过程。

大数据服务战略过程首先是定义市场。大数据服务与企业向客户提供的产品或者服务一样，都能为客户带来价值，定义市场的目的是找到大数据服务的服务对象。大数据服务的对象可能位于企业外部，也可能位于企业内部。如果大数据服务的对象在企业外部，则需要明确大数据服务的客户群体、供应品、渠道、价格等。例如，大数据服务是为了支持企业完成面向某客户群的营销，那么大数据服务应当首先从全部客户中挑选出服务的子集，向这些群体推广。大数据服务也可能服务于企业内部，帮助企业提升战略管理能力和运营效率，此时应当明确大数据服务的业务驱动力、适用场景、所需数据源、关键实现活动等。

当市场和供应品定义完成后，需要完成战略资产的开发。战略资产主要是指大数据服务资产，而大数据服务资产是企业提取、转换、控制、交付的一种能力。与服务资产相对的是客户资产，企业应当借助大数据服务提升客户资产，将服务资产和客户资产紧密地结合起来。

当大数据战略资产开发完成后，需要在企业的生产和运营阶段，完成战略的执行。当在战略执行的过程中发现问题后，需要评估大数据服务执行的效果。如果执行效果不满意，可以优化和完善分析模型，调整模型参数，然后再次投入到企业的生产与运营过程中。由于大数据服务是一个探索发现的过程，企业需要不断调整分析方法和手段，优化和完善模

型和算法，从大数据中找寻和发现价值。

3.1.4 大数据服务战略组织

大数据服务是企业的一种无形资产，同样可以作为产品进行销售，将给予大数据构建的产品称为大数据产品。

如果是大数据服务类似于某种业务，那么大数据产品则是在对业务进行组合，增加销售属性后形成的。

大数据产品的形成不是一蹴而就的，与其他虚拟产品（服务）一样，需要经过市场调研、产品定义、产品开发、产品销售、客户服务、产品评价的全过程，而大数据产品经理需要关注整个过程。

由于大数据产品经理在大数据运营过程中的重要作用，企业应当在大数据服务战略阶段就要优先考虑大数据产品经理的权责问题。

大数据产品经理的任职要求主要包括：

（1）能够发现大数据中潜在的价值。企业可以要求大数据产品经理具有一定的行业背景，比如通信行业的大数据人才要具备 3 年以上的通信行业从业背景，掌握信息通信相关知识。

（2）优先考虑具有大数据分析和挖掘经验的人士。由于大数据更多地依赖不断地探索尝试后才能增强对于大数据价值的认识，因此具有大数据分析和挖掘经验的人士对于大数据的潜在价值有更好的直觉。

（3）具备保护组织商业机密和个人隐私不被侵犯的知识，掌握相关法律法规。由于数据很容易复制和传播，一旦泄露，就像泼出去的水，难以收回，因此大数据产品经理应当掌握保护个人和组织隐私的方法和手段。

大数据产品经理的工作职责包括：

（1）负责大数据产品的市场调研、分析，并完成大数据产品定义；

（2）负责完成大数据产品的成本效益分析。由于大数据产品是企业的一种虚拟产品，在资产负债表中以无形资产（软件）的形式存在，因此难以估量大数据产品的成本；

（3）根据大数据产品的推广效果和使用反馈，对大数据产品进行调整和完善。由于大数据产品的用户将企业提供的大数据服务作为一种数据源来集成和分析，因此可能会对大数据服务提出新的要求，比如在现有大数据服务的基础上增加新的数据项。

一般而言，企业内部的大数据服务在符合法律规范的条件下，都可以打包成大数据产品对外销售。由于大数据通常包括客户属性和客户行为记录，很可能会涉及企业商业秘密或者个人隐私，因此大数据产品经理应该着重考虑数据的安全性和合规性问题。

3.2 大数据服务设计方法论：方法比努力更重要

首先分析大数据可能具备的能力，然后再分析问题域的特点，最后结合大数据能力与问题域特点，形成大数据服务需求。

面向操作的事务型应用的需求由业务人员根据企业生产和经营需要提出，而大数据服务的需求则是由数据分析人员在探索发现过程中逐步确定的。前者具有稳定性和确定性，而后者则具有偶然性和不确定性；前者通过满足企业生产和经营需求达到降本增效的目的，而后者则是提升组织的决策管理能力。

大数据服务需求分析以组织大数据和待解决的管理问题为输入，输出大数据服务的需求。首先对大数据能力进行评估，分析大数据具备的能力；其次是对待解决的问题进行分析，找出问题域中的决策点；最后是将大数据能力与待解决问题进行综合分析，找出借助大数据解决问题的思路和方法。

3.2.1 大数据服务设计原则

面向操作的事务处理系统的设计原则主要包括可靠性、可用性、可伸缩性、高性能以及安全性 5 个方面。面向决策的分析处理系统与面向操作的事务处理系统的特点不同，因此设计的原则也存在非常大的差异。

在系统的可靠性方面，为了满足日常的生产经营需要，通常需要对事务处理系统在网络、计算、应用和数据层面进行可靠性设计，比如网络双路由、服务器集群、冗余磁盘阵列（RAID）、中间件集群、数据库集群等，原理是通过冗余的资源换取可靠性。对于面向决策的分析处理系统而言，嵌入生产流程之中的实时性大数据服务和本身就是一种事务型应用（这就是 ODS 存在的原因），因此对可靠性的要求与面向操作的事务处理系统一样，但是分析处理系统对于可靠性的要求相对要低一些，原因是分析系统内部的数据本来就是

历史数据，是外部数据源的“备份”，同时分析系统通常是“正确性”大于“快速性”，分析时长可以以分钟、小时甚至天来计，对于实时性响应的要求不高，要求决策信息要准确，要能够发现规律。

此外，面向决策的分析处理系统对于系统的可用性要求也没有面向操作的事务处理系统那么高（基于 ODS 的 OLAP 应用例外）。在可伸缩性方面，由于分析处理系统的数据源可能会源源不断地增加历史数据，因此要求系统要有良好的可伸缩性和扩展性，企业在进行大数据服务架构设计时，一定要保证基础设施资源具有良好的横向扩展性。

在系统性能方面，对于即时查询这样的应用，用户需要能够快速提供查询结果（比如 5 秒之内）；对于统计报表这样的应用，要求分析系统的响应性能越快越好，这样决策者就可以更快地拿到统计结果，可以通过构建中间表和采用内存数据库的方式实现；对于像数据挖掘这样的分析型应用，其特点是从大量的数据中找出数据之间的联系，对于系统性能的要求相对较低。

在安全性方面，不同的分析型应用有不同的安全级别要求，需要区别对待，制定不同的安全管理制度。比如对于涉及个人或者企业隐私的大数据服务，需要通过采用加密、审批等手段来保证数据不被非法获取。

除了以上 5 个方面，数据的准确性和完整性是做出正确的决策的必要条件，因此要考虑大数据质量保证问题，由于数据分析结果更多地体现为统计特征，因此数据质量没有事务处理系统要求那么高。

3.2.2 大数据服务需求分析方法

软件需求是成果交付的依据，也是系统实现的前提。大数据服务的特点就是难于事前提出明确的需求，而是在模糊需求的基础上逐步确定需求的。

可以将大数据服务需求分析的方法称为“距离拉近法”，就是将企业大数据可能具备的能力与要解决的决策问题之间的距离逐步拉近。分解开来就是首先分析大数据具备的能力，然后分析待解决问题的特点，最后通过分析实验确定大数据服务的需求。

企业的业务活动可以分为战略、战术、操作三个层次，无论在哪个层次都有决策活动，随着企业运用大数据进行决策的事情越来越多，如何对大数据服务进行有效管理就会越来越重要。企业业务活动分为“执行”和“决策”两类，“执行”就是按照预先设定的规则完成任务，而“决策”则是解决是否应当“执行”以及如何“执行”的问题。

大数据与大数据服务需求之间的关系类似于“鸡生蛋、蛋生鸡”这样的问题，其实谁先谁后并不是最重要的。为了解决特定问题，有的情况下需要分析有没有数据可以解决这一问题，而有的时候则是具备了数据基础，然后再去看看它到底能够解决企业的什么问题。

大数据服务需求分析的方法如图 3-2-1 所示。

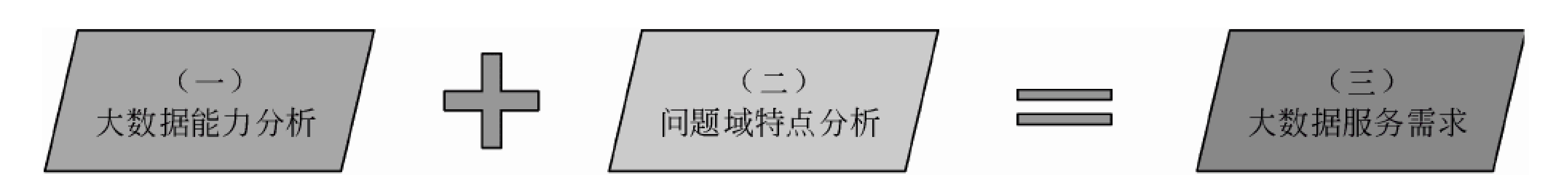


图 3-2-1　大数据服务需求分析方法

从图 3-2-1 可以看出，可以首先分析大数据可能具备的能力，然后再分析问题域的特点，最后结合大数据能力与问题域特点，形成大数据服务需求。

企业通过汇集来自不同来源的数据，形成了一个大的数据资源池，这时就有了形成大数据分析的基础，应当初步查看这些数据具备什么能力。以企业获取的电信大数据为例，通过对数据项的分析，发现数据包含用户、网络、应用三个方面的信息，也就具备了三个方面的能力。

企业要清楚希望借助大数据解决什么问题，比如借助大数据提升市场营销能力，提升网络规划设计能力或者提升客户服务能力等。以电信大数据为例，电信运营商希望借助大数据提高无线网络规划能力，即通过运用大数据，确定哪些地方应当新建或者扩容基站，新建或者扩容多大规模等。

有了对大数据能力的分析和待解决问题的分析，将两者相结合，就可以形成大数据服务需求。例如，通过对电信大数据能力的分析，发现其具备用户、网络以及应用三个方面的能力，通过对无线网络规划设计决策需求的分析，发现无线网络规划设计的关键是确定哪些区域应当新建或者扩容，新建或者扩容的规模有多大。那么，如果要解决无线网络规划问题，可以从时间、区域、网络类型等维度出发，结合用户价值和应用价值，确定那些需求量大但无线网络资源不足的区域，根据需求和现有资源之间的差值计算新建或者扩容的规模。

1. 大数据能力分析

大数据是对客户世界和人类社会属性和行为的记录，由多种数据源汇聚而成的大数据意味着多种不同的能力，可以帮助企业更好地把握客户偏好，可以帮助企业更好地完成资

源规划，等等。

对大数据的能力进行分析是确定大数据服务需求的第一步。如果抛开大数据可以解决的问题不谈，可以按照下面的方法分析大数据具备的能力。

企业要管理的对象包括客户、产品、渠道、资源、人力资源、财务、资产等，对于任何企业，应当在有限资源的前提下满足客户需求，实现最佳的成本效益，因此对于大数据的能力分析，可以以客户（用户）、产品（商品）、渠道、资源、人力资源、财务、资产等数据为抓手，对数据进行整合，达到从多个角度发现大数据能力的目的。

比如，企业取得了来自电信运营商、银行、电子商务公司的数据，通过对以上数据进行整合，发现大数据具备把握客户的能力，具体包括客户消费能力、客户购物偏好、客户通信行为等。企业收集的数据越全面，越能够准确地把握客户。

下面以电信大数据为例，分析电信大数据具备的能力。电信大数据主要包括用户通话记录和用户上网记录，通过对电信大数据的分析，发现其具备客户、网络、应用三个方面的能力。

从客户角度看，客户发现用户使用的终端能力和行为特征。终端能力包括是否可以上网，是否支持 4G 网络等。行为特征包括通话时间、上网时间、通话时长、上网时长、使用终端、使用的网络、访问的应用、访问地等。

从网络角度看，可以发现用户在通信网络的访问路径、访问时长、访问流量、源 IP 地址、目标 IP 地址等。

从应用的角度看，可以发现访问该应用类型、访问流量，根据应用类型，如新闻、音乐、视频、电子商务等进行归类分析，根据流量判断该应用的活跃度和价值。

可见，通过对大数据的分析，可以发现大数据具备的能力，这为解决企业决策中遇到的问题做好了准备工作。

2. 问题域分析

通过对不同来源的数据进行整合后，可以发现大数据具备的能力，这为人们利用大数据解决现实问题打下了基础。但是，如果不能识别待解决问题的特点，也无法利用大数据解决现实问题。

问题域分析往往需要具备待解决问题域的专业知识，对于待解决问题所在领域的专业知识掌握得越好，越有助于快速地发现和解决存在的问题。

下面从企业战略、战术、操作三个层面，分析企业如何利用大数据解决决策问题。

在战略层面，大数据主要支持企业的高层管理人员完成决策，完成的任务包括投资战略规划、市场战略规划、人才战略规划、信息化战略规划等，大数据更多的是为战略规划提供数据参考，比如企业要投资，大数据可以给出投资对象的投资风险数据、企业财务风险数据等。

在战术层面，大数据主要是支持企业的中层管理人员完成决策，完成的任务包括市场营销能力提升、产品销售能力提升、客户服务能力提升、客户体验能力提升、运营成本节约等。在这个层次，大数据主要是帮助企业中层管理人员做管理决策。比如企业产品销售经理要提高产品销售能力，可以利用大数据分析客户的属性和偏好，然后为客户提供所需的产品和服务。

在操作层面，大数据主要是支持企业的基础人员完成决策。当企业使用信息系统完成操作层次的任务时，大数据也可以嵌入操作流程中，辅助完成决策任务。比如，当客户通过银行的网上银行贷款时，银行可以利用大数据预先进行信用评估，确定客户的贷款额度，这样客户就可以在没有人工参与的情况下完成贷款，提升了客户感知，也提高了银行的办事效率。大数据在企业操作层次的应用，其实就是用机器智能代替人脑的过程。

3. 需求分析

当完成对大数据能力和问题域的分析后，就可以发现问题以及借助大数据解决问题的方法。大数据服务的作用就是能够为解决问题提供决策支持，因此大数据服务的需求就是借助大数据解决问题的方法。

电信运营商具有移动用户上网的使用记录，这些大量的记录就是通信大数据，通过对来自 CRM 系统的用户数据、来自网络运营支撑系统的用户上网行为数据以及来自外部的辅助数据的采集与整合，就可以形成关于用户价值、应用价值、用户网络访问路径等大数据能力。

同时，问题域中存在的核心问题是如何提升移动用户上网速度，其特点是解决移动用户因跨地域和跨电信运营商网络而引起的网络速度下降问题，其方法就是借助大数据发现某地域的移动用户访问高价值应用是否存在跨地域和跨电信运营商网络问题，如果是，则可以结合大数据对该地域移动用户平均价值的分析结果，决定是否需要新增 CDN 节点。高价值应用为访问流量排名靠前的应用，移动用户价值的判断标准可以人为设置，比如以平均 ARPU 为 100 元/月为标杆，大于这个值的为高价值应用。

通过以上分析可以看出，可以根据大数据能力和问题域分析确定大数据服务需求，然

后根据大数据服务需求来设计和开发大数据服务，最后借助大数据服务实现决策支持。

4. 需求管理

大数据服务需求是大数据服务设计、转换和运营的输入以及大数据服务改进的基线，它决定了大数据服务的范围，影响到大数据服务占用的成本、资源、时间等，因此对于大数据服务需求的管理非常重要。

在企业的生产和运营阶段，主要以市场为导向，以客户为中心，将企业生产和运营的需求作为构建信息系统的输入。在企业决策阶段，即大数据服务阶段，主要是以大数据作为资源输入，构建满足企业生产和运营决策的大数据服务。可见，数据的形成和数据的使用是一个闭环的反馈过程，达到“取之于企业，用之于企业”的目的。

随着数据源的不断丰富，大数据服务的功能越来越强，数量越来越多，因此只有通过对大数据服务需求的管理，才能够保证大数据服务之间不会产生重叠和交叉现象。由于大数据服务的需求就是企业业务活动中的决策需求，因此对大数据服务的需求管理可以采用分层分类的方法，与企业业务过程框架的分层分类管理方法是一致的，这样可以避免因大数据服务需求不断增多难以管理的问题。

大数据服务需求管理的方法为：基于大数据，初步分析大数据可能会形成的能力以及可能解决的问题，要解决的问题一定在企业业务过程框架中，因此要在企业业务过程框架中找到大数据服务的问题。以移动用户上网记录大数据为例，通过初步分析发现，移动用户上网记录大数据可以解决无线网络规划问题，那么就可以将这个大数据服务放到 SIP（战略/基础设施/产品管理）域，然后再进入 SIP 域内部，发现该大数据服务属于“基础设施生命周期管理”过程组和“资源开发与管理”过程组的交叉部分，因此可以将大数据服务先放到这个位置，以后再根据对大数据服务更加深入的理解，对该过程块进行细分，使得大数据服务的分类更加准确。

通过将大数据服务放置到企业业务过程框架之中，实现对大数据服务需求的有效管理以及大数据服务的能力共享。大数据服务与企业业务活动的紧密结合，可以帮助用户准确定位大数据服务的发力点。

3.2.3 大数据服务开发方法

亚马逊 CTO Vogels 在 Cebit 上发表的主题演讲称：“大数据不仅仅是分析，它是关于

整个流程的。当你思考大数据的解决方案问题时，要考虑所有的步骤：收集、存储、组织、分析和共享。”

可见，对于大数据来说，不能仅仅关注结果，而应关注整个过程，需要认真对待每个过程环节，以便最大限度地发挥大数据的价值。

为了便于分析，将应用分为两类：操作型应用和分析型应用。操作型应用用于支撑企业的生产和运营。分析型应用用于支撑企业在战略、建设、生产和运营过程中的决策。

从需求与数据的关系看，操作型应用是先有需求后有数据的，大数据服务属于分析型应用，其特点是先有数据后有需求，因此两种不同类型的应用在设计方法上也是不同的。

操作型应用的主要目标是支撑企业的生产与运营，提高企业管理能力，因此操作型应用的需求是业务驱动的。分析型应用的主要目标是支撑企业战略、建设、生产、运营过程中的决策，是决策驱动的。

尽管操作型应用与分析型应用的目标有很大的不同，但是从软件工程的角度看，两者也是具备许多共同点的。比如在满足需要变化方面，操作型应用主要是要满足业务需求的变化，而分析型应用主要是要满足决策需求的变化，两种都需要通过快速迭代来适应这一变化。

大数据服务的特点之一是通过“过去”来预测“未来”，通过对数据的不断整合分析和挖掘找出事物之间的联系。这个过程就是一个不断试错的过程，通过不断地调整模型和算法，发现数据背后隐藏的规律，为决策提供更丰富、更全面的参考依据。

敏捷开发强调沟通、反馈、简单、勇气和谦逊，属于小步快跑的开发模式，这种方法可以快速发现并修正错误，降低了软件工程风险。

大数据服务能够满足的需求主要在于数据源是否丰富，大数据服务的能力主要取决于数据质量的好坏。如果在现有数据源的基础上增加新的数据源，则可能会提高决策能力，如果对现有数据源继续进行清洗，提高数据质量，那么大数据服务的决策能力则会得到提升。

考虑到数据规模、范围、质量等对于大数据服务能力的影响以及大数据服务主要用于决策参考，因此大数据服务设计更适合于采用敏捷开发方法。当然，敏捷开发并不代表前期规划并不重要，大数据服务同样需要从全局和长远进行考虑，企业需要真正把大数据当作核心资产来管理，根据探索反馈实现设计的持续改进，最终达到为企业提供决策支持的目标。

3.3　大数据服务架构设计：在平衡中实现完美

大数据服务运营框架从业务角度出发，体现业务到数据的互动过程，大数据服务应用框架从能力角度出发，体现了大数据的管理过程。

大数据服务的实现要经过数据采集、存储、管理、分析、治理直至实现各种应用的周期，为了对大数据服务进行有效管理，需要将大数据服务从数据获取到需求实现的过程进行细分，细分为几个相互区别又相互联系的子部分，最终形成大数据服务的框架体系。

第一步是解决数据源问题。数据源是大数据服务形成的基础，数据可以来自于企业内部和企业外部，可以来自于不同的行业，不同形式的媒体、不同的地理位置以及采用不同的时间段。

第二步是完成不同来源数据的收集和存储。在这个阶段，主要考虑如何将所有数据源收集和存放起来，要保证将来自不同数据源的数据存放到不同数据仓库之中。

第三步是根据大数据服务的要求，对数据进行加工，使其达到满足大数据分析的目标。大数据服务需求会随着人们对大数据认识的逐步深入而不断发生变化，因此对于大数据的加工完善是一个持续渐进的过程。

第四步是面向待解决的问题域，基于大数据进行数据建模、数据分析并形成不同的主题应用。比如提升收入和客户体验的应用、降低生产与运营成本的应用等。

此外，大数据服务全过程中要考虑隐私、安全以及合规性问题，保证利用大数据资产的同时，满足隐私保护、法律规范等社会性要求。

大数据服务管理是一个系统化工程，为了对其进行有效的管理，需要构建一个系统化的框架体系。

3.3.1　大数据服务运营框架设计

通信、信息、交通等技术和工具的发展，加速了全球化进程，全球产业分工合作已是既定事实。《世界是平的》做了形象的描述：“在印度 24/7 的呼叫中心，你会发现电脑操作系统是微软的，芯片是英特尔的，电话是朗讯的，空调是凯利的，饮用水是可口可乐的，

甚至公司 90%的股份都是美国投资者的。”

历史发展进化的历程说明了一个道理：适者生存、优胜劣汰。为了适应当前的市场环境，满足客户需要，企业需要借助先进的技术工具提高竞争力。

然而，尽管业界发明创造了许多先进的方法与工具，但是业务与技术之间的仍然存在天然的鸿沟：业务人员侧重于市场客户，其对于业务的理解更接近于现实生活，而技术人员偏重于逻辑思维，更多地考虑在技术框架下如何将业务需求转化为技术实现。

为了缩小业务与技术的鸿沟，笔者采用业务与技术分离的架构形式，将大数据运营框架分为活动层、应用层和大数据层，如图 3-3-1 所示。

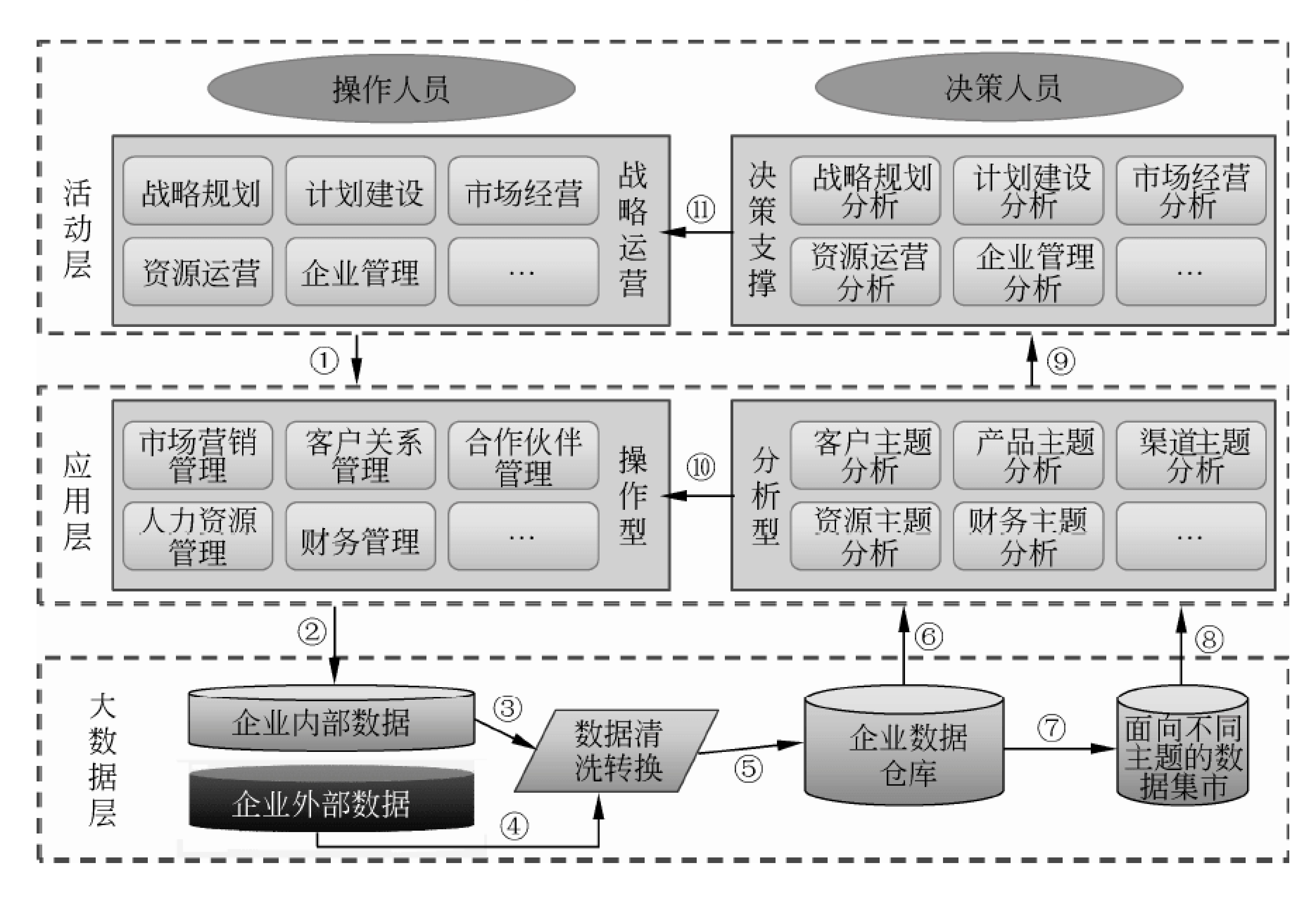


图 3-3-1　大数据运营整体框架

从图 3-3-1 可以看出，活动层属于业务层面，其参与方主要包括市场、客服等面向客户的角色以及人力资源、财务、资产等面向内部管理的角色，活动层的主要目标是完成企业的战略与运营活动，而大数据则是支持战略与运营活动的重要工具和手段。

大数据层属于技术层面，其参与方主要包括信息系统分析、设计、开发实施等相关的角色，大数据层的主要目标是对来自于企业内部和外部的信息与数据进行收集、加工、存储、组织、分析、分享等。

应用层介于活动层和大数据层之间，是活动层和大数据层之间的桥梁和纽带。从活动

层看，应用层体现了业务人员和管理人员的能力需求，从大数据层看，应用层体现了技术人员需要向业务人员和管理人员交付的 IT 能力。

在现实中，活动层是企业在生产经营过程中所做的事情，业务活动不一定依靠信息系统来实现，比如校园现场宣传活动。应用层是抽象的 IT 能力集合，对于活动层其承载业务需求，对于大数据层，其承载数据能力。

3.3.2 大数据服务应用框架设计

大数据服务运营框架从业务角度出发，完成整体框架设计，体现业务到数据的互动过程。大数据服务应用框架从能力角度出发，体现了大数据的管理过程。大数据服务应用框架分为数据采集、数据存储、数据分析、数据治理、数据应用几个部分，如图 3-3-2 所示。

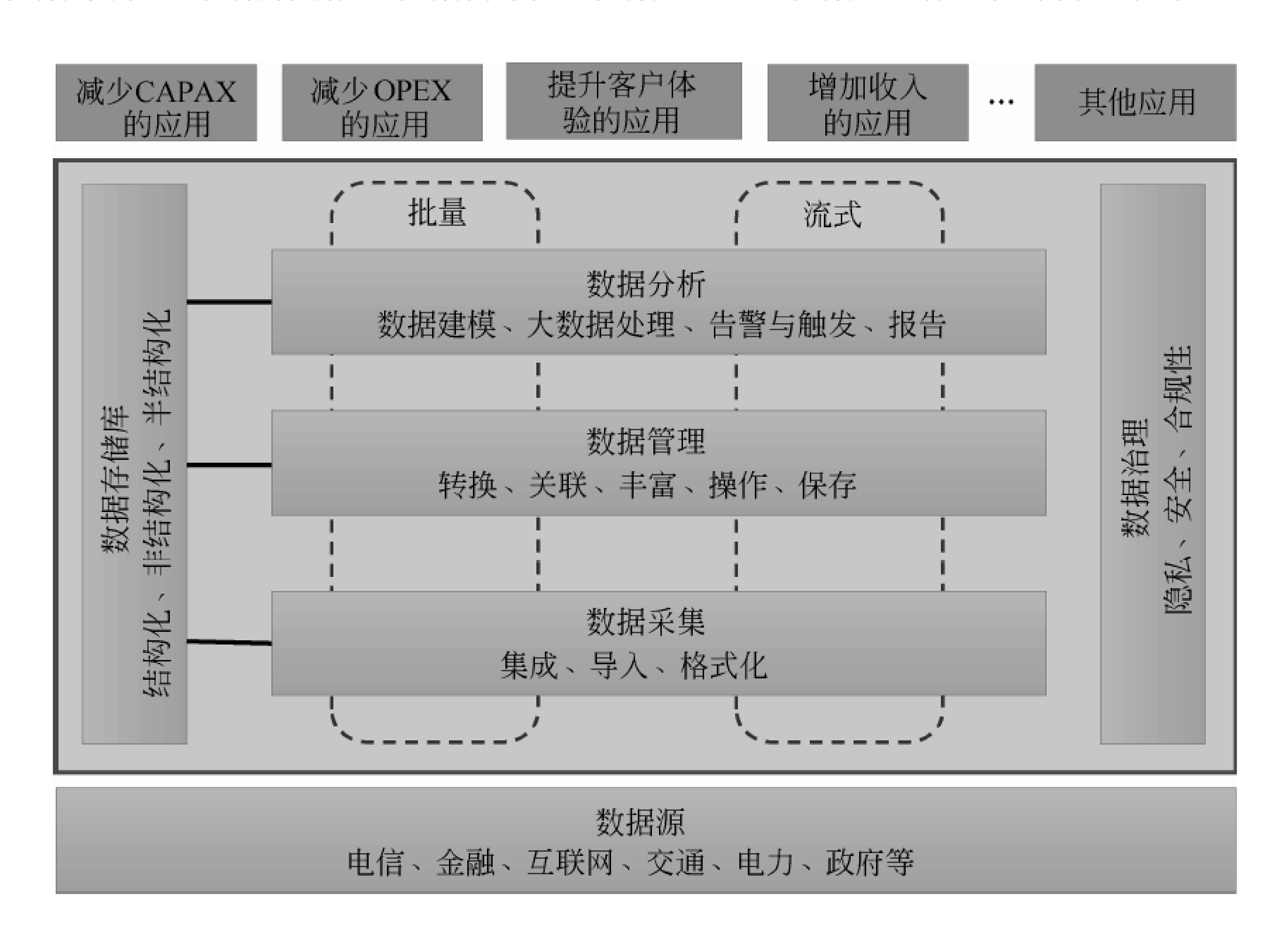


图 3-3-2　大数据服务框架体系模型

从图 3-3-2 可以看出，大数据服务应用框架分为数据源、数据采集、数据存储、数据管理、数据分析、数据治理几个部分，这些部分相互配合，形成各种不同类型的大数据分析应用。

3.3.3 大数据服务数据源

大数据记录了个人和企业在社会活动中的行为，而这些行为必然存在于为个人或者企业提供服务的组织之中。一般来讲，大数据服务的数据源是越多越好，充足的数据能够保证企业决策时考虑得更全面。

随着社会的发展进步，整个社会被分为提供不同服务的企业，比如提供信息通信服务的电信运营商、提供存贷款业务的银行、提供保险服务的保险公司、提供在线购物服务的电子商务公司、提供公共交通服务的公交公司、提供供电服务的电力公司、提供公共管理服务的政府等。

人们在生产生活中，不但是服务的使用方同时也是服务的提供方，正所谓“我为人民服务，人民为我服务”。在个人享受社会不同组织提供的服务的同时，也将个人行为记录在信息系统之中。

除了个人信息和行为会记录在服务提供方的信息系统之中，也会将企业行为记录在企业自身的信息系统之中或者与企业交互的组织的信息系统之中，比如，企业纳税记录会记录在税务局的信息系统之中，企业的采购记录会作为供应商的销售数据记录下来，企业与合作伙伴的交互记录会记录在合作伙伴的信息系统之中，等等。

可见，大数据服务所需的数据源存在于各个专业化组织之中，比如电信、交通、电力、金融、政府、互联网等。将这些分散在不同类型的服务提供商之中的数据进行整合，就会形成一个关于个人或者企业的行为全景图，数据越全面，全景图就越完善，大数据服务也就有了更加强大的数据支持。

3.3.4 大数据服务数据采集

数据采集过程包括集成、导入、格式化。

数据采集过程中首先集成来自不同来源的数据。数据集成要考虑存储架构、采集方式、接口方式、采集周期等。

在存储架构方面，可以考虑在数据源侧设置数据暂存区（Staging Area），也可以考虑在采集平台侧设置暂存区。靠根据数据量和累计速度来设置合理大小的数据暂存区，防止数据溢出。

在存取方式方面，可以根据应用的需要采用不同的存取方式。采集方式包括单个采集和批量采集两种类型，对于数据量小、时效性要求高的应用，可以采用单个采集的方式，当数据形成后可以立即同步到数据仓库。比如用于审计的操作日志，可以采用单个采集的方式，当操作日志产生后就实时地同步到数据仓库。对于文件多而且实时性要求相对较低的数据，可以等文件数达到一定规模或者达到一定的时间周期后，批量采集或者推送到数据仓库。

在接口方式方面，对于批量采集的数据，可以考虑采用 FTP 方式，对于单个采集的数据，可以采用 API 或者 Web Services 接口的方式。

在采集周期方面，通常是采集周期越短，数据的实时性越高，数据分析的结果越及时。企业可以根据应用的需要设置不同的采集周期，要考虑数据暂存区能否满足要求。

在数据导入方面，根据数据规模大小分为三种导入类型。

第一种是数据量大而且需要导入数据定义的场景，比如数据定义包括索引、分区等，可以考虑采用大文件导入方式，这样可以保证数据源的完整性。

第二种是对于数据源结构简单、导入文件多、规模大的数据，可以采用批量文件导入的方式，这样可以看到导入过程中产生的错误，并及时纠正，保证数据导入的质量。

最后一种是对于数据量小的单个文件，比如某些代码表、配置文件等，可以通过数据导入工具逐个导入，这种方式比较简单灵活。

数据采集阶段的数据规范化工作非常重要，因为数据分析必须基于一个统一的标准，而多种数据源就某一个数据通常会存在形成和内容上的不同。比如在 A 数据源中，日期格式以“年-月-日”的形式存储，而 B 数据源中以“月-日-年”的形式存储，因此需要将这两种数据源中的格式进行统一。也有的字段存储的数据类型不一样，比如在 A 数据源中，年龄字段以字符串格式存放，而 B 数据源中以整型格式存放，需要将两个字段统一为一种数据类型。还有的数据在不同数据源中存放的内容不一样，但是表达的是同一个意思。比如 A 数据源中的“性别”以 M 和 F 代表“男”和“女”，而 B 数据源中的“性别”则是用 1 代表“男”，而用 0 代表“女”，因此需要实现两种数据源“性别”在语义上的统一。

不同数据源在同一数据上存在差异的原因是信息系统设计时并没有考虑到其他信息系统或者不同的应用提供商并没有遵循共同的编码规范。

3.3.5 大数据服务存储库

大数据服务的数据源不但来自归属于不同行业的组织之中，而且其类型还具有多样性

（Variety）特征。

多样性指的是大数据服务不仅包括例如姓名、年龄这样的结构化数据，还包括歌曲、电影这样的非结构化数据，此外网页、邮件这样的数据介于结构化和非结构化之间，属于半结构化数据，也是大数据服务的重要数据源。

结构化数据来源于业务需求，系统分析员将需求中静态的“名词”提取出来并进行抽象，作为数据库表结构设计的依据。比如设计一个学籍管理系统，通过分析发现“张三”、“李四”等学生具有姓名、年龄、所属院系、所选课程、课程分数等属性，于是系统分析员将这些属性选取出来并设计一个“学生”类，那么“学生”表结构就相当于一个模板，可以将“张三”、“李四”等学生的姓名、年龄、班级等结构化数据存储到数据表中。由于数据表是二维的，借助关系型数据库的 SQL，可以从多个维度对结构化数据进行查询统计。

与结构化数据相对的是非结构化数据。顾名思义，非结构化数据是不可以提取字段并定义属性的，只能以图片、语音、视频的媒体形式存在。虽然非结构化数据不像结构化数据那样能够进行统计分析，但是并不代表非结构化数据没有价值。非结构化数据可以以多媒体的形式存在，生动形象地反馈信息，因此可以从非结构化数据中采集有价值的信息，并将这些采集的信息转化为结构化数据，通过对非结构化数据的“理解”来发现其中隐藏的价值。

介于结构化数据和非结构化数据之间的是半结构化数据。半结构化数据的结构和内容混合在一起，例如电子邮件、网页等。从半结构化数据中同样可以抽取出许多有价值的数据，比如电子邮件中可以采集到发件人、收件人、标题等，通过对邮件的收发地址、频率、主题等进行分析，可以形成以电子邮件为通信媒介的社交网络。

企业可以根据应用的要求、数据的规模、数据的类型等维度进行分析和设计，选择不同的存储架构。

对于数据规模大、数据结构简单、对查询效率要求高的应用，可以采用 Hadoop/HBase 这样的分布式存储架构。由于 Hadoop/HBase 存储架构采用键值存储结构，具有良好的可扩展性，因此可以通过增加基础设施资源来提高查询效率，系统整体性能随着集群规模的增大而线性增长。

对于需要关联多个数据模型才能实现的分析型应用，则可以考虑采用关系型数据库作为存储库。对于以邮件、文档、录音、录像等文件形式存在的非结构化数据，可以采用网络连接式存储（Network Attached Storage，NAS）架构，对于存取频率高、单次存取数据量小的结构化数据，具有明确数据类型和数据长度，可以考虑采用存储区域网络（Storage

Area Network，SAN）存储架构。对于以文件为存取单位的非结构化数据，则适合采用网络连接式存储（Network Attached Storage，NAS）存储架构。通常情况下，存储架构采用 SAN 和 NAS 混合的形式。

SAN 和 NAS 属于“主机+磁盘阵列”的系统架构，在大数据时代，随着数据量的不断增加，企业越来越采用“单机+硬盘”组成的系统架构。这种架构适合于需要批量数据处理的分析型应用，并且对单个应用设备的能力要求不高，可以有效地利用旧低端设备，快速地实现横向资源扩展。

3.3.6 大数据服务数据管理

数据管理过程主要包括数据转换、数据关联、数据丰富、数据操作以及数据保持。

数据转换就是将数据从一种形式变换为另一种形式，通过形式的变化，使得数据更便于分析利用。比如在数据采集阶段导入的原始数据，需要将其从字符串类型转换为浮点型，这样可便于对该数据项进行求和。另外，也可能因为数据格式问题进行数据转换，比如原始数据为网页这样的半结构化数据，为了能够搜索到网页中的数据，往往需要将网页中的关键数据提取出来并做成标签，再把标签作为检索项，这样检索时就没有必要检索整个网页了，通过这样达到提高检索效率的目的。

数据关联是按照需要，借助关联属性将多个分散的数据源关联在一起，就像用一根绳子将多个数据串接起来一样，目的是方便定位所需数据，同时便于从多个维度进行数据统计。比如，身份证号码、手机号码、终端设备号、网络编码等可以作为数据关联的外键，也可以根据分析需要构建多个数据表，以实现数据的关联。

范式原则可以提高操作型数据模型对业务需求响应的灵活性，减少数据冗余，分析型数据模型则希望通过数据关联形成面向多个主题的数据模型，面向主题的数据模型更加接近于用户需求，便于多维度地分析和展现数据。

数据丰富也是为了满足业务需求而对数据进行的完善，比如有一个学生，如果只知道她的姓名、性别信息，不知道她的生日信息，就不能知道这个学生的年龄。再比如知道承载某个应用的 IP 地址及其产生的数据流量，如果再知道这个 IP 地址对应的 URL，就可以知道这个 URL 对应的数据流量。

数据操作就是操作数据，包括数据联合、去重、排序、过滤、分组等，通过数据操作，实现数据的关联与组合，便于从不同视角对数据进行查看和统计。

数据保持要考虑数据的存储策略，包括分散存储还是集中存储，采用原始表存储还是中间表存储，基于内存存储还是外存存储，存储周期多长，按月存储还是按年存储等。

数据保持对于大数据存储管理非常重要，制定数据保持策略的参数包括数据价值高低、数据活跃度、存储策略、法规要求等。企业可以根据需要定义数据的价值，比如客户的身份证号码、出生年月、家庭关系、教育经历、偏好等数据的价值比较高并且这些属性比较稳定，而对于客户购买历史、支付历史等数据则相对要低一些，因此难以作为预测未来的数据基础。

数据活跃度也是数据保持的一个重要指标，通常来说，数据存取频率高的数据的价值要高，如果数据超出一定的时间（比如一年）没有被存取，可以考虑将其转移到低价值的存储空间，以便提高数据的存取效率。

企业可以根据数据的特点和用途，对数据的生命周期进行定义，比如电信运营商超过一年的账单数据可以存放到二级磁盘阵列，通常将面向客户查询的详单数据存储在一级磁盘阵列。

当然，有些存储策略并不是企业根据生产经营需要制定的，而是基于国家法律法规的要求制定的，比如，政府要求银行对储户的交易数据至少保留 6 年，要求电信运营商对用户的通话记录至少保留 3 年，等等。

数据的存储周期越长，用于数据分析的样本数据就越多，越容易从长期的数据变化中发现规律。企业需要综合考虑数据分析的实际需求、数据存储成本、数据管理成本等因素，实现企业成本效益的最大化。

3.3.7 大数据分析

数据分析过程包括数据建模、大数据处理、告警与触发、报告等。

从数据处理的实时性要求角度看，大数据分析可以分为批量和流式两种数据处理方式。批量处理主要适合于实时性要求不高的分析型应用，而流式处理主要适用于实时性要求高的在线分析应用。

批量处理方式主要适用于大规模离线数据的分析处理，比如企业周期性统计报表，可以采样批量处理方式。对企业大规模历史生产经营数据进行批量处理，分析结果可以用于制定企业发展战略，对于分析结果的实时性要求不高。

流式处理方式有许多应用场景，比如客户浏览网页时，企业可以实施实时的产品推介

或者广告投放，当客户使用手机访问应用时，可以根据客户的位置和访问的应用，向客户推送附近商家最新的促销信息。社会关注热点分析也是流式处理的一种典型应用，可以基于搜索大数据，实时展示社会关注热点。

价值创造是大数据分析的目标，数据建模、大数据处理、策略执行以及分析结果展示过程，对体现大数据的价值都有非常重要的作用。在大数据处理阶段，采用批量处理还是流式处理方式，取决于应用的要求。

3.3.8 大数据治理

数据治理包括隐私、安全、合规性三个方面。

大数据价值创造的前提和基础是企业自身的数据以及全社会开放的数据，当数据开放为社会带来好处的同时，也同时引起了隐私侵犯问题。隐私是社会赋予个人或者企业的权利，隐私权受到法律的保护，因此，企业在利用大数据的同时，要首先考虑大数据应用是否会侵犯他人或者组织的隐私。

企业可以多种方式来解决隐私触犯问题。

对于企业向内部人员提供的大数据服务，可以通过数据权限保证隐私数据不被非法获取，如果企业内部用户具有获取隐私数据的权限，要进行数据使用行为的记录和跟踪。例如，电信运营商拥有公众客户电话号码、银行卡号、家庭住址等隐私数据，可以采用授权的方式控制数据使用对象和数据使用范围，系统应当能够自动记录数据操作行为，实时进行数据使用行为审计，发现可疑数据使用行为后，计算采取措施，关闭或者暂停用户的数据访问权限。

企业对外提供大数据服务具有更大的风险，就如同覆水难收，因此，企业需要考虑更好的隐私保护方式。以电信运营商为例，用户的姓名和电话号码是不能泄露的，如果这些数据被营销机构所掌控，用户可能会经常接到骚扰电话或者短信，电信运营商可以为外部企业提供电话号码的伪码数据，企业如果想与电信运营商提供的名单客户沟通，还需要借助电信运营商提供的伪码翻译服务，这样就解决了客户真实的电话号码外泄问题。

企业大数据治理的另一个难点是数据安全问题。应用分为事务型和分析型两种，大数据服务属于分析型应用，相对于事务型应用，大数据服务安全治理具有自身的特点。

可以将安全控制分为应用和网络传输两个层面。应用层安全控制包括用户安全管理和

信息安全管理，用户安全管理的目标是让系统设定的用户访问应用，并对认证用户进行授权，保证用户访问所需的资源。信息安全管理的目标是保证信息不被非法获取，通常采用对信息加密的方式实现。在网络传输层实施安全控制的目标是控制进入网络的通道，通过安全控制策略来阻止或者进行网络访问。

事务型应用是创造数据的源头，产生的数据可以分为基础数据和交易数据两类。交易数据是在每个事务处理之后产生的，比如网页浏览记录、订单数据。与交易数据相比，基础数据的内容变化频度要低，比如客户的姓名、年龄、身份证号等数据，相比于订单数据，其生命周期要长，数据的安全性要求更高。许多企业的事务型应用暴露到互联网，因此网络安全风险高。

分析型应用的数据基础是事务型应用产生的数据，通常要经过采集、转换、装载、分析、展示或者对外提供的过程。根据大数据服务用途的不同，可以分为企业内部使用和对外提供两种类型。相比于对外提供的大数据服务，在企业内部范围使用的大数据服务风险要低得多。如果从数据的规模来看待数据风险，分析型应用比事务型应用依赖的数据规模要大得多，因此，一旦出现数据泄露，分析型风险要大得多。

从系统架构的角度看操作型数据和分析型数据，操作型数据通常以“主机+磁盘阵列”的集群方式存放在磁盘阵列中，而分析型数据则通常以“主机+磁盘”的集群方式分散存放数据仓库的磁盘上，由于采用批量处理方式，集群内部主机之间往往没有实施安全控制，同时，由于数据规模大，为了提高数据处理效率，一般不会对数据进行加密。

企业需要根据事务型应用和分析型应用的不同特点制定不同的数据安全管理机制。

3.3.9 大数据服务应用

对于企业来说，大数据服务的目标可以归结为“降本增效”4 个字。企业可以借助大数据服务做精准化营销，将企业的产品有效地传递给有此需求的用户，在为客户创造价值的同时增加企业收入。

企业也可以借助大数据服务掌握客户偏好，更好地为客户提供服务，提升客户感知水平，虽然提升客户服务体验并没有直接为企业带来收入，但是通过这种方式提升了企业在客户心中的形象，使得客户获取企业服务更加便捷、高效，客户也因此更喜欢购买企业的产品，从而增加了企业的收入。

除了增加企业的收入，企业还可以借助大数据服务降低成本。从费用支出的类型角度看，成本消耗主要分为属于投资建设的 CAPAX 投资和属于业务运营的 OPEX 投资两部分，因此企业可以借助大数据服务降低这两部分投资。比如在降低 CAPAX 投资方面，可以以用户价值为中心进行资源的建设，避免因为靠“假设”、“猜想”而造成投资浪费。在降低 OPEX 投资方面，企业可以借助大数据服务来发现企业流程中存在的问题，通过流程优化来提高运营效率，从而降低企业的整体运营成本。

3.4 大数据服务模型设计：默默无闻的贤内助

行成于思而毁于随，面向操作的数据模型侧重对“行”的支持，而面向分析的数据模型则侧重对“思”的支持。

为了便于看到数据从形成、聚集、整合、使用的全过程，从面向操作的数据模型和面向分析的数据模型两个阶段分别设计。

从面向操作的数据模型到面向决策的数据模型的转化过程如图 3-4-1 所示。

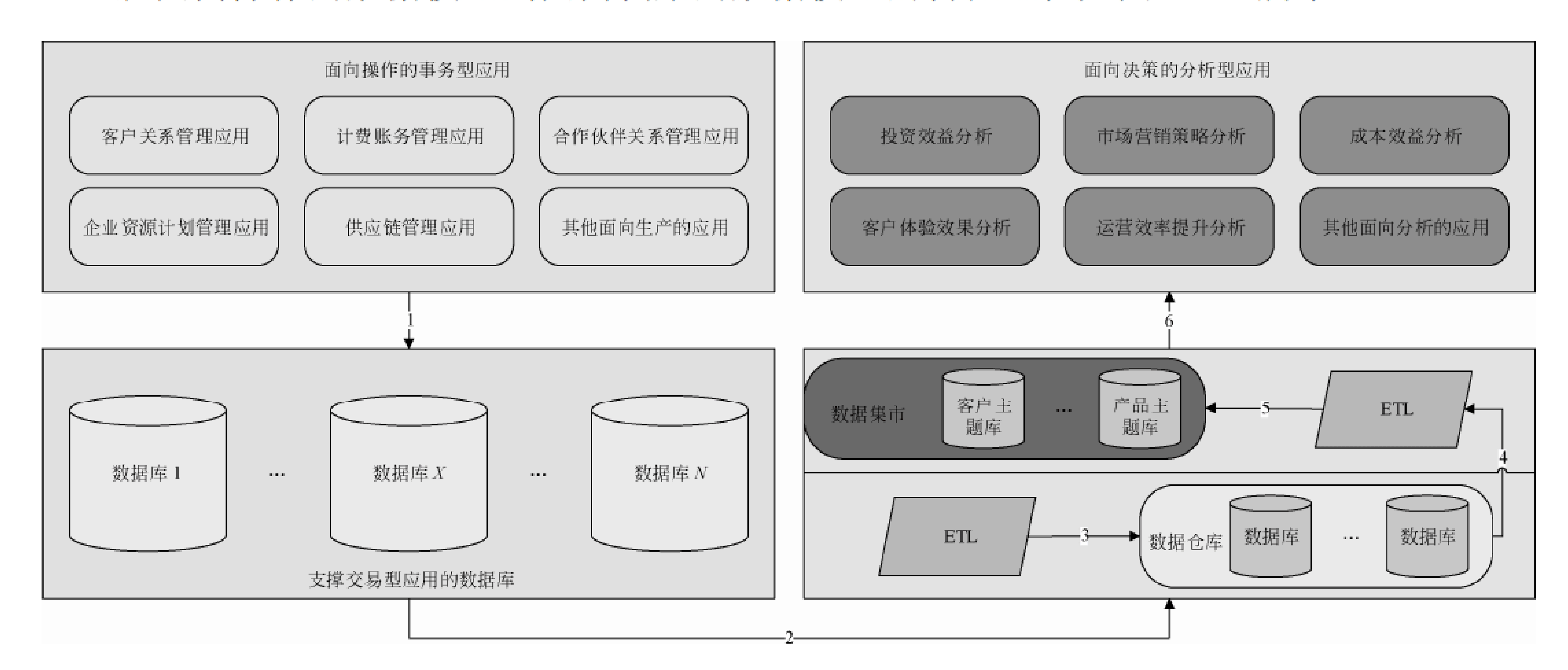


图 3-4-1　数据从面向操作环境到面向分析的环境的转化

从图 3-4-1 可以看出，操作型数据来源于事务型应用，这些数据会存储在支撑事务型应用的数据库之中，来自不同来源的数据会汇聚到数据仓库，然后通过 ETL 等工具和手段，形成面向不同主题的数据集市。

以电信运营商为例，在面向操作的事务型应用会记录客户以及客户的订购数据，如果再加上客户的业务使用数据，就可以形成一个以客户为中心的、360° 的数据集合，通过数据的分类聚合，可以帮助企业全面地看到客户从咨询、订购、支付、使用、申告、投诉、建议的全过程，从而达到帮助企业决策的目的。

按照数据服务的目的，将数据模型分为面向操作的数据模型和面向分析的数据模型。面向操作的数据模型主要支撑企业完成数据的增加、删除、修改、查询等操作，帮助企业完成建设、生产、运营以及内部管理等任务。而面向分析的数据模型主要完成对不同主题、不同维度统计分析功能的支持。

可见，面向操作的数据模型侧重对“行”的支持，而面向分析的数据模型则侧重对“思”的支持。“行成于思而毁于随”，没有行动则思考没有素材，没有深刻的思考则行动很可能会偏离方向，企业需要统一“行”和“思”。

从数据模型对企业应用的支撑层次，将数据模型分为战略模型、战术模型和操作模型，如图 3-4-2 所示。

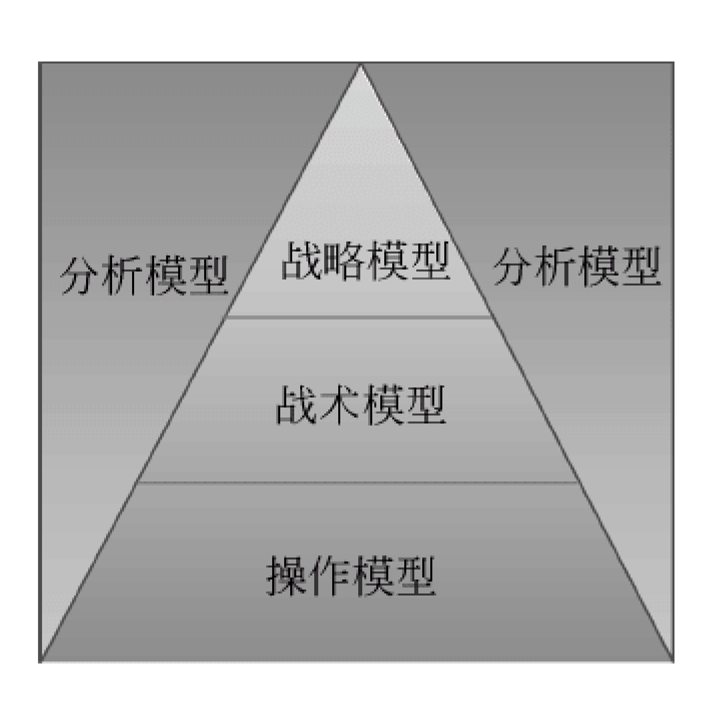


图 3-4-2　战略、战术、操作三个层次的数据模型

在战略层面，面向分析的数据模型支撑企业高层管理人员完成战略制定工作，包括工厂选址、渠道选址、兼并与收购规划、投融资规划、环境影响分析、非常规资金预算等。

在战术层面，面向分析的数据模型支撑企业的中层干部完成管理工作，包括市场营销计划、销售计划、客户服务计划、人力资源计划、财务预算等。

在操作层面，面向分析的数据模型服务于企业的日常生产经营活动，包括个人贷款授信、生产进度安排、库存控制、维护计划、质量控制等。

面向操作数据模型与面向分析的数据模型好比汽车发动机和油门/刹车之间的关系。面向操作的数据模型好比汽车发动机，保证汽车的正常运行、转弯，而面向分析的数据模型

则好比司机根据路况确定踩油门或者刹车一样，两种必须配合起来，才会保障企业的业务活动有序进行。

3.4.1 面向操作的数据模型设计

在企业的生产经营过程中，客户关系管理系统、供应链管理系统、企业资源计划系统等产生了各种数据，而承载以上数据的模型就是面向操作的数据模型。

古语云："水能载舟亦能覆舟"，如果将面向操作的数据比作"水"，那么面向分析的数据就像"舟"，如果没有"水"，那么"舟"就没有了基础，可见操作型数据模型在大数据服务中的重要地位。

操作型数据模型承载的数据是以操作对象为单位的，比如客户、产品、渠道、订单等。操作型数据通常存放在事务处理系统的数据库之中，由于这些数据是企业可控的，因此成为企业进行数据分析的重要基础。

为了更加清晰地看到数据从产生到整合利用的过程，本节对面向操作的主要数据模型进行分析与设计，包括产品数据模型、客户数据模型、渠道数据模型、资源数据模型、供应商/合作伙伴数据模型、人力资源数据模型、财务数据模型以及资产数据模型。

1. 产品数据模型设计

产品是企业的核心载体，集中反映了市场需求和资源供给。一方面，产品反映了企业对市场需求的满足，包括产品销售区域、产品满足的客户群、产品营销渠道、产品客户服务渠道等。另一方面，产品反映了企业自身资源的供给能力。一个企业的资源总是有限的，企业满足的市场需求是基于企业自身资源的供给能力的，包括人、财、物等各种类型的资源。

可见，产品在企业中具有核心作用，下面就从设计产品数据模型说起。

从产品全生命周期看，包括产品设计、产品测试、产品营销、产品销售、产品评价、产品退出的全过程。

在产品设计阶段，要考虑产品的构成、资费、品牌、营销区域、营销渠道、面向客户群等因素。

在产品测试阶段，通过对产品的内部测试，保证产品满足设计阶段的功能性和非功能性要求，保证产品可用。企业也会通过将产品投放到特定市场区域的方式进行产品测试。

在产品营销阶段，主要包括产品营销活动计划的制订、营销活动的执行效果评价等，

目标是将产品信息传送给有产品需求的人群。

在产品销售阶段，主要是保证产品有效地交付给客户，保证产品交付的效率和效果。

在产品评价阶段，主要是对投放市场一定时间周期的产品进行效果评估，根据产品评价结果调整市场营销、销售以及客户服务策略。要将不符合市场需求或者已经过期的产品退出市场。

根据以上分析，企业产品数据模型如图 3-4-3 所示。

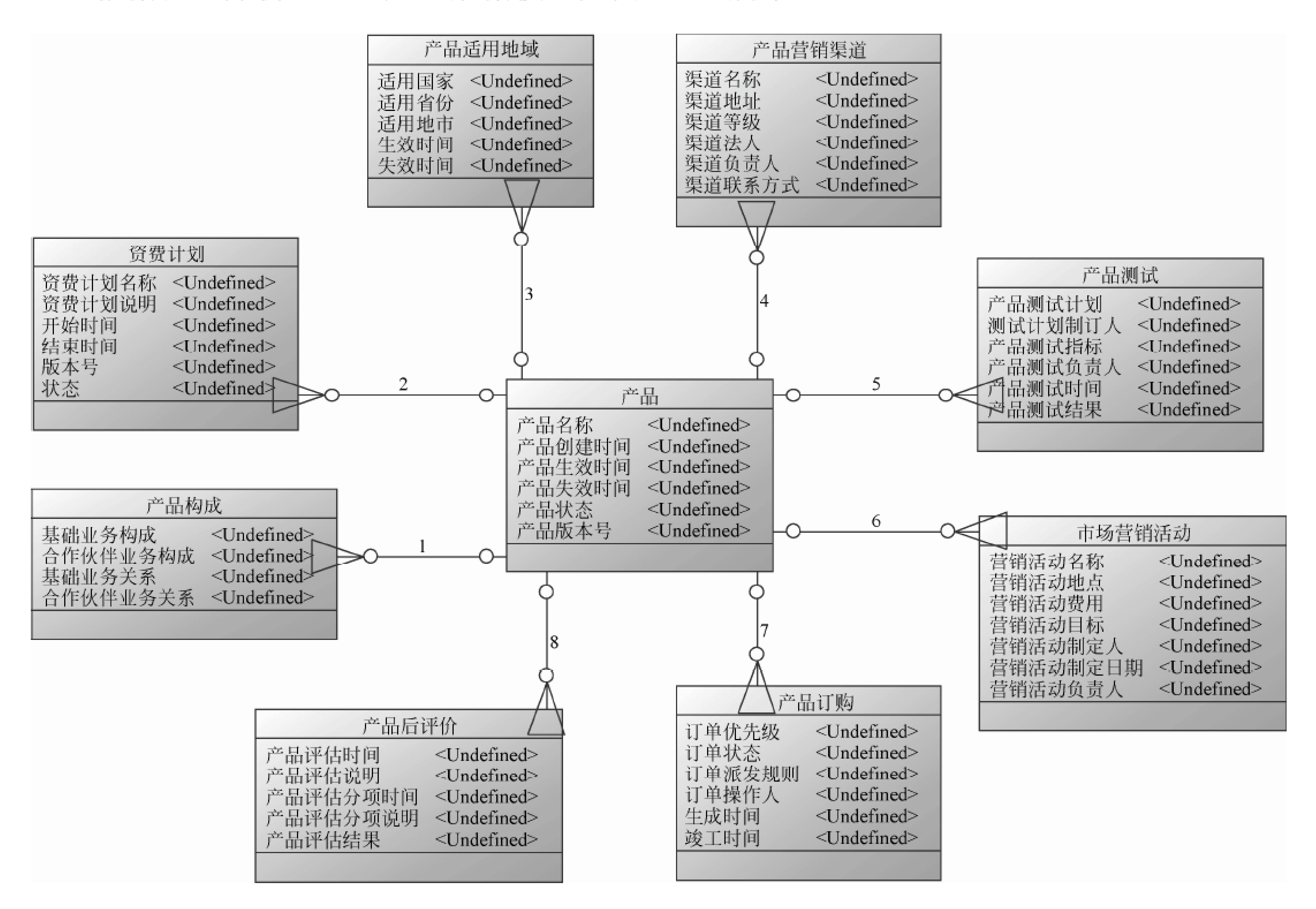


图 3-4-3　企业产品数据模型示例

从图 3-4-3 可以看出，产品数据模型关乎企业从市场营销、产品销售、资源供给的各个方面，为了模型设计的灵活性，上图的模型中没有直接体现产品与资源的对应关系，而是通过业务来衔接产品和资源的关系的。

2．客户数据模型设计

客户不仅包括已经使用企业产品或服务的客户，也包括潜在客户。随着社会生产能力

的提升，社会商品越来越多，产品同质化趋势越来越明显，企业借助互联网，满足个性化的客户需求，从长尾中获利。

客户个性化需求越来越多，使得企业认识到只有生产满足市场需求的产品，才能够降低企业库存压力，防止生产过剩或不足现象发生。因此，企业采用了以客户为中心的管理思想，对于客户全生命周期进行管理，包括对细分市场，精确定位客户群等。企业客户概念模型如图 3-4-4 所示。

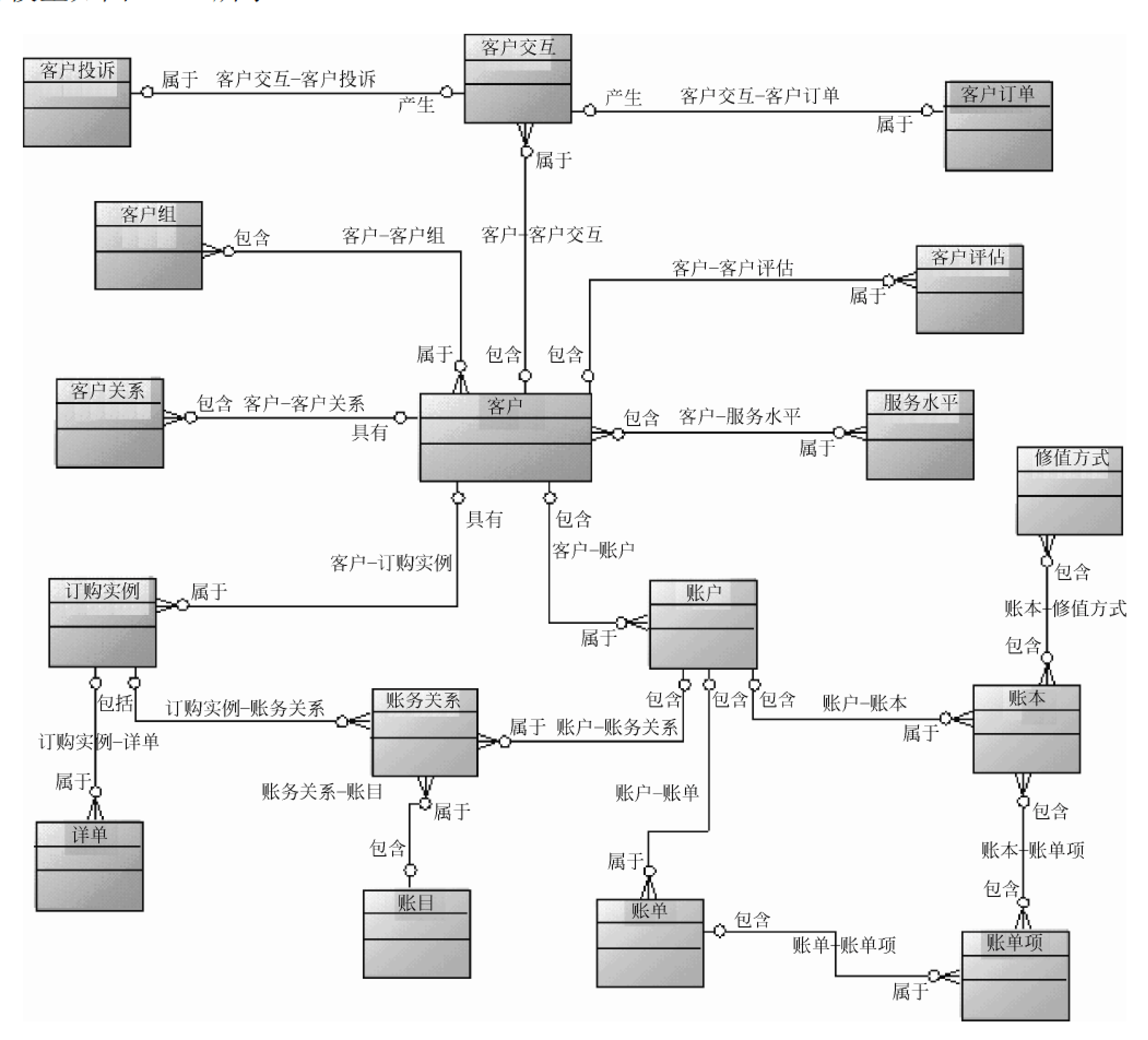


图 3-4-4　企业客户概念模型示例

图 3-4-4 中的模型称为三户模型，已经长期应用于电信运营商的业务支撑系统之中。三户分别代表客户、用户和账户。客户体现了社会域的信息，用户也称为订购实例，体现了业务域的信息，而账户则体现了资金域的信息。

3. 渠道数据模型设计

渠道是企业将产品和服务交付给客户的一种手段。渠道可以整合各种产品和服务，在产品日益同质化的今天，渠道在社会中的重要性越来越突出。渠道在产品和服务的提供方和消费方之间的作用如图 3-4-5 所示。

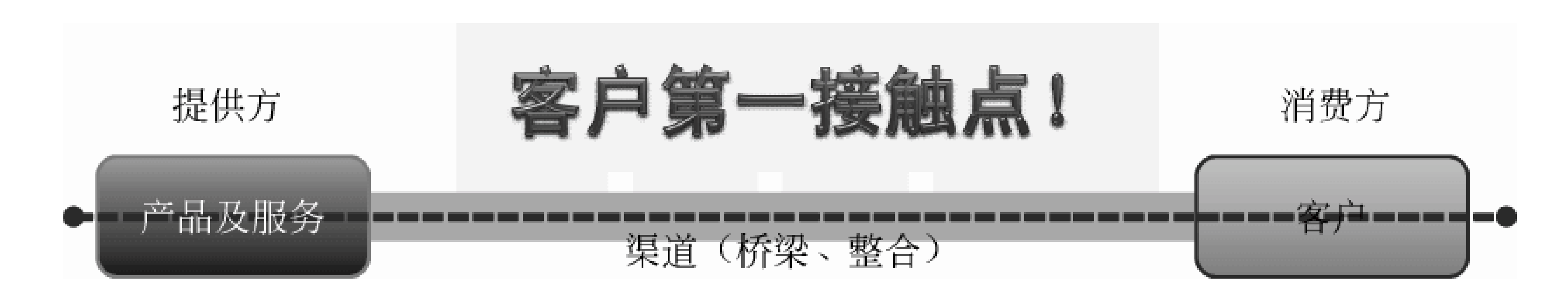


图 3-4-5 渠道是连接客户与产品/服务的桥梁和纽带

按产权归属，可以将渠道分为自有渠道和社会渠道；按照存在形态，可以将渠道分为实体渠道和电子渠道。企业的渠道服务体系如图 3-4-6 所示。

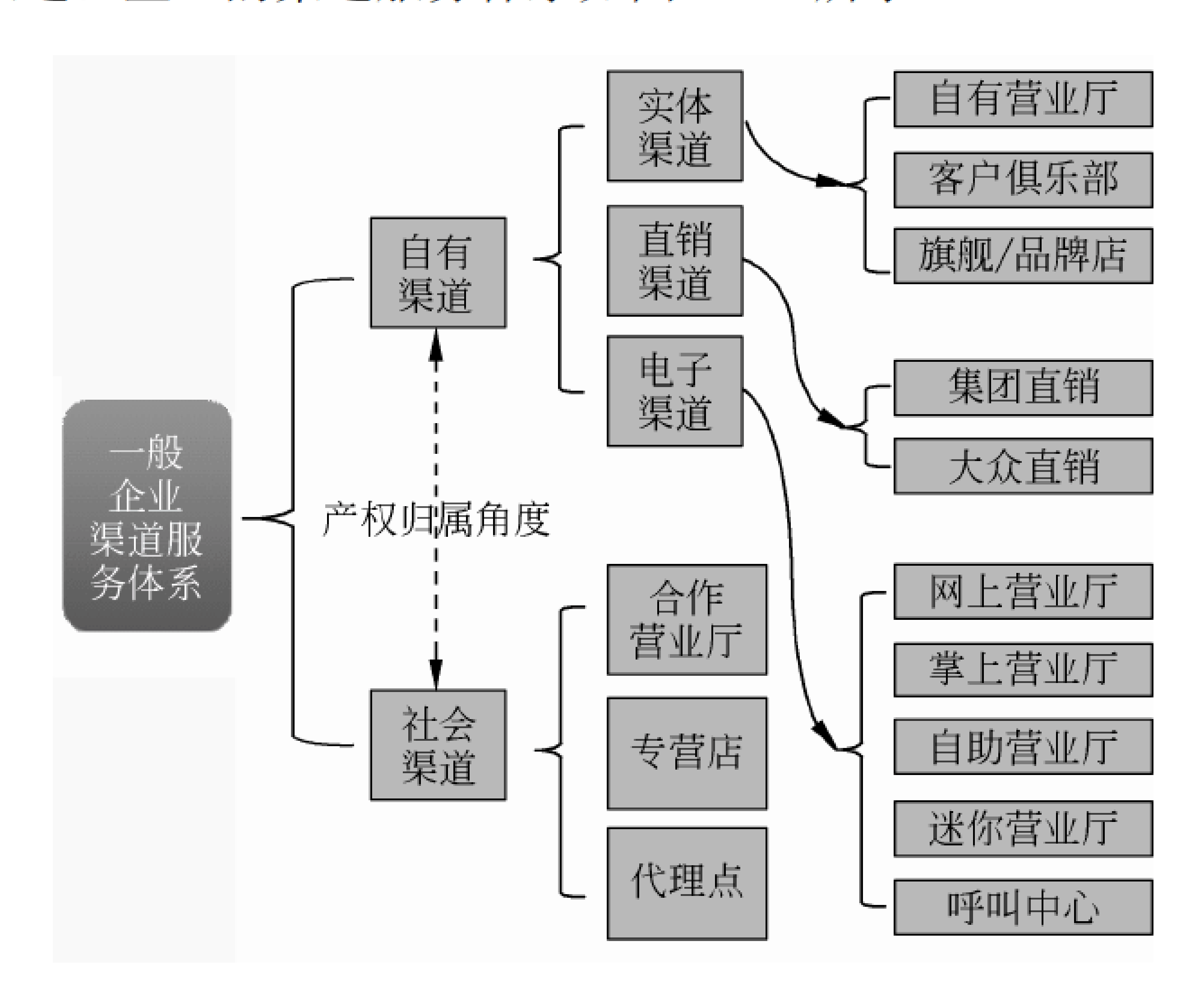


图 3-4-6 企业渠道服务体系示例

电子渠道可以降低产品和服务交付成本，提供客户获取产品和服务的便捷性，在互联网发达的今天，电子渠道在渠道体系中的占比越来越高，但是由于实体渠道能够获得实物体验，电子渠道和实体渠道还是一种互补关系，像服装、家居等需要现场体验的商品，还需要借助实体渠道销售。企业可以采用线上到线下（Online to Offline，O2O）模式，将实

体渠道和电子渠道结合起来，发挥两种渠道各自的优势，实现两者的有效协同。

企业渠道的概念模型如图 3-4-7 所示。

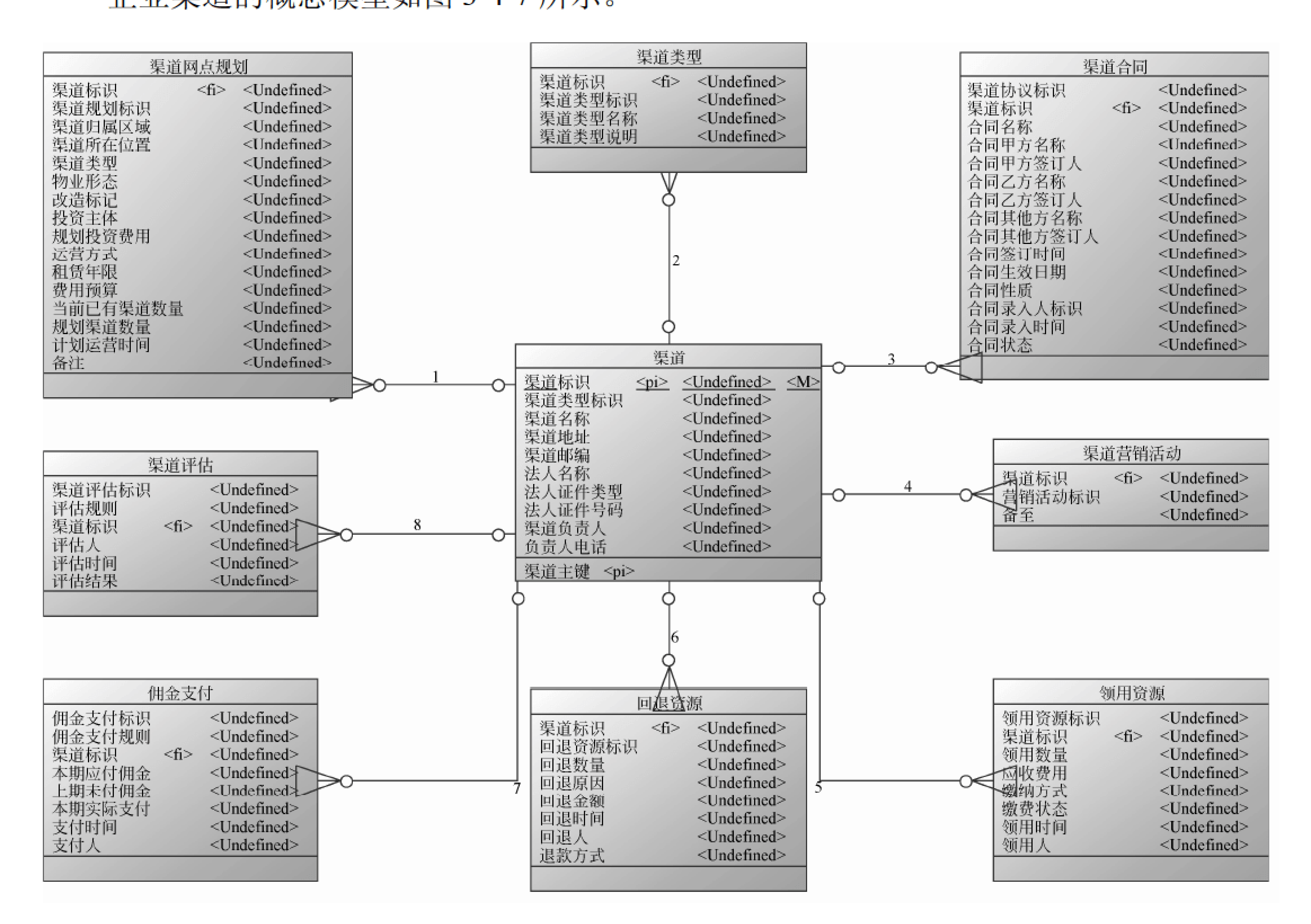


图 3-4-7 企业渠道概念模型示例

4. 市场营销数据模型设计

企业的市场营销、销售以及客户服务活动属于动态行为，反映了企业与客户交互的过程，对市场营销进行模型设计的目的是捕捉交易过程，为客户提供更好的服务。

市场营销的目的是吸引客户并促进销售。一般包括营销战略与规划、营销活动策划、营销区域、接触/机会、竞争对手、销售统计以及销售渠道。

企业的市场营销概念模型如图 3-4-8 所示。

5. 资源数据模型设计

资源的定义很广，在这里是支撑生产与价值创造的物质，比如网络资源、货币资源、

IT 资源、业务资源等。

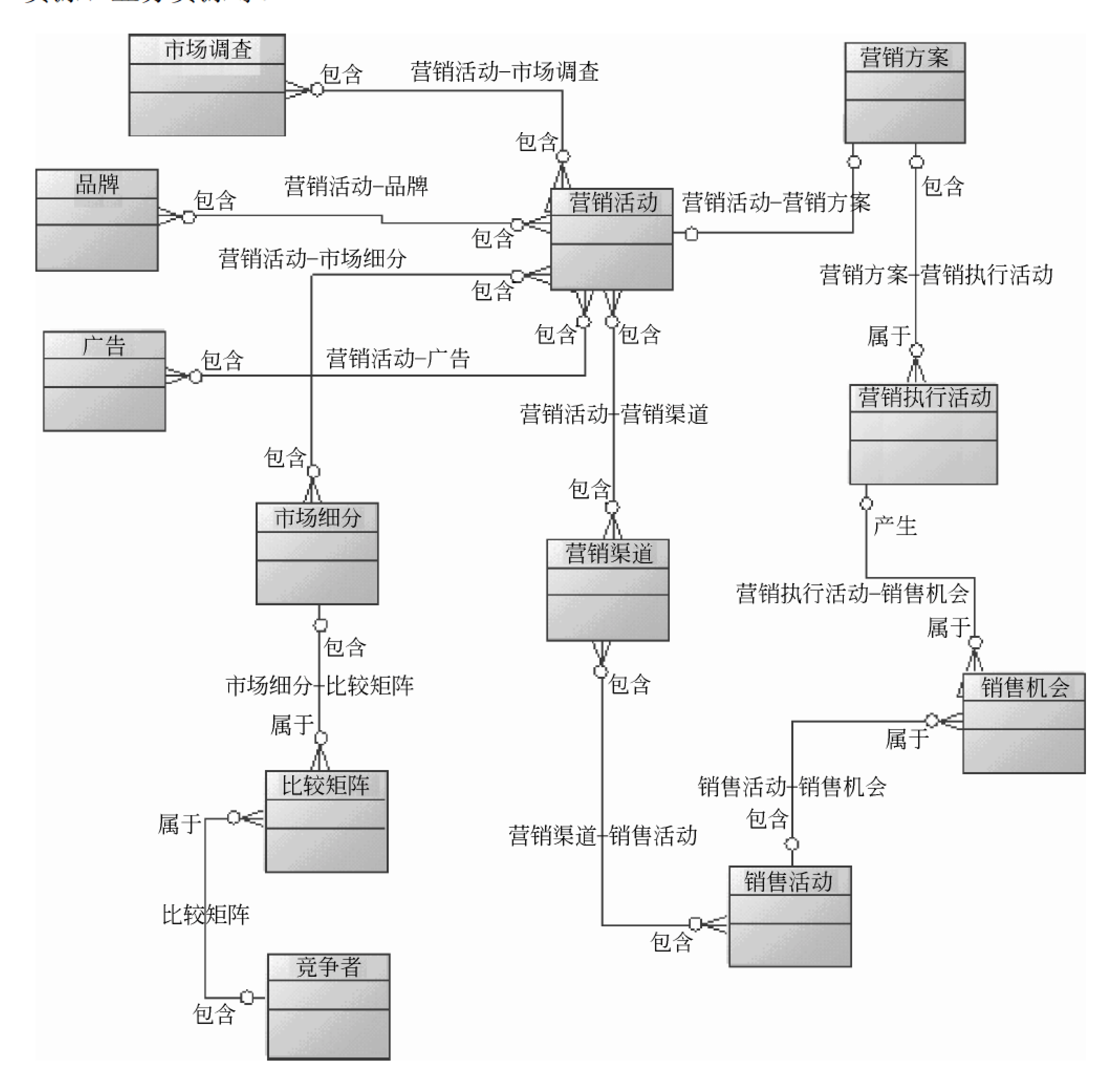


图 3-4-8　企业市场营销概念模型示例

同样，资源也是区分不同行业的重要依据。比如，电信行业的资源主要是指信息通信网络资源，网络才是电信运营企业提供信息通信服务的根本；金融行业的资源主要是指货币资源，因为货币才是银行、证券、保险等金融企业的根本，没有货币资源，这些行业就失去了存在的基础；互联网行业与电信行业类似，资源主要是指 IT 资源，比如交换机、路由器、服务器、存储、中间件、数据库等，如果没有 IT 资源，互联网企业就无法为用户提供搜索、新闻、电子商务等线上服务。

此外，业务资源也是很重要的资源之一，因为业务资源是直接面向营销的，比如卡、号、终端、礼品、票据等，这些业务资源在完成产品和服务的交付中也起到非常重要的作用。

资源分为物理资源和逻辑资源。物理资源是指有形的实体，逻辑资源是指无形实体。物理资源包括人们能够看到的机架、机框、板卡、插槽、端口等，逻辑资源包括 IP 地址、逻辑端口号、各种软件等。

资源主要面向生产和市场经营，更多的是体现其实物属性。与资源相对的是资产，它主要体现其价值属性，是企业进行成本核算的重要依据。通信资源数据模型如图 3-4-9 所示。

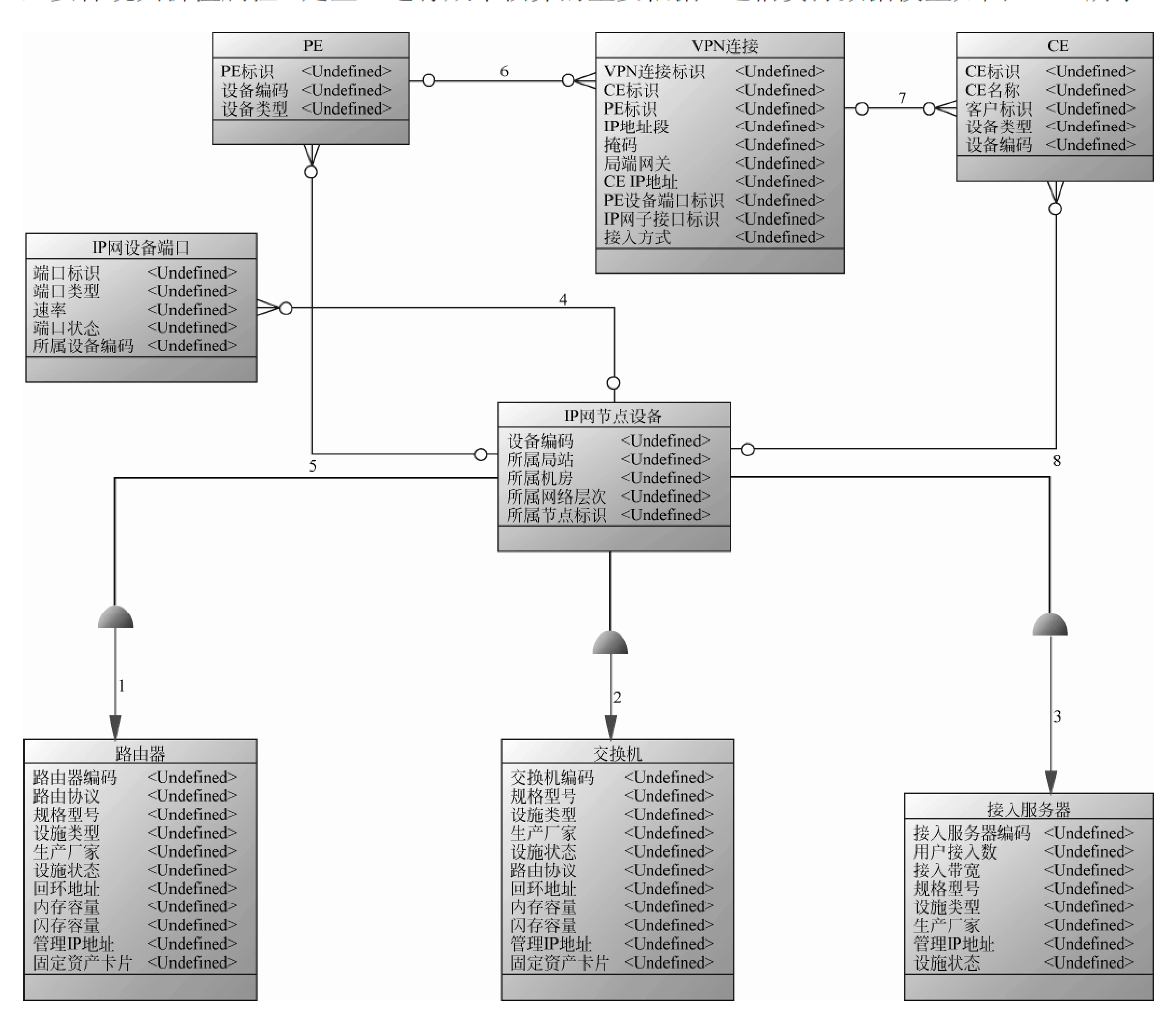


图 3-4-9　企业资源数据模型（以通信资源为例）

以上为不同行业中的基础支撑性资源，除了基础支撑性资源之外，随着产品的日益同

质化，市场竞争日益激烈，客户、品牌等无形的资源在企业中的地位越来越重要。

6．供应商/合作伙伴数据模型设计

广义的合作伙伴是指一切与企业合作的对象，包括供应商、分销商、服务提供商、内容提供商等。在这里，合作伙伴特指那些提供产品、服务和内容的供应商，这些供应商提供的产品、服务和内容并不是用于企业的基础设施建设的，而是用于企业产品和服务的构建的。

企业合作伙伴管理数据模型应当能够实现企业对产品与服务提供商/内容提供商的信息管理，企业与合作伙伴的接触管理、合同管理、违约管理、培训管理、评估管理等。企业合作伙伴数据模型如图 3-4-10 所示。

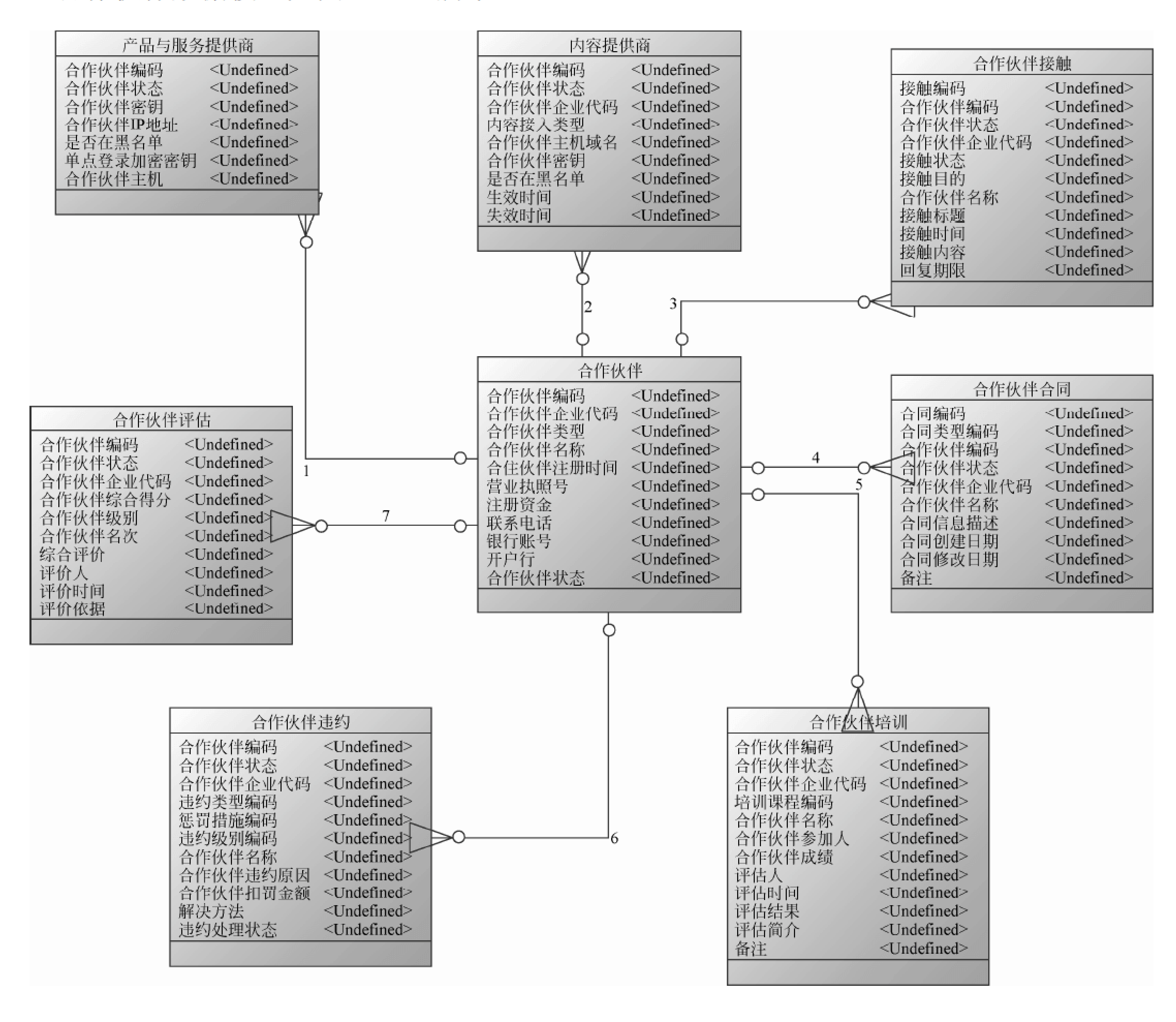


图 3-4-10　企业供应商/合作伙伴概念模型示例

7. 人力资源数据模型设计

员工是企业创造价值的主体。企业应当采用全生命周期管理的思维方式来实现对人力资源的管理，包括员工招聘、员工考勤、薪酬福利管理、绩效考核、培训考试等过程环境。企业人力资源数据模型如图 3-4-11 所示。

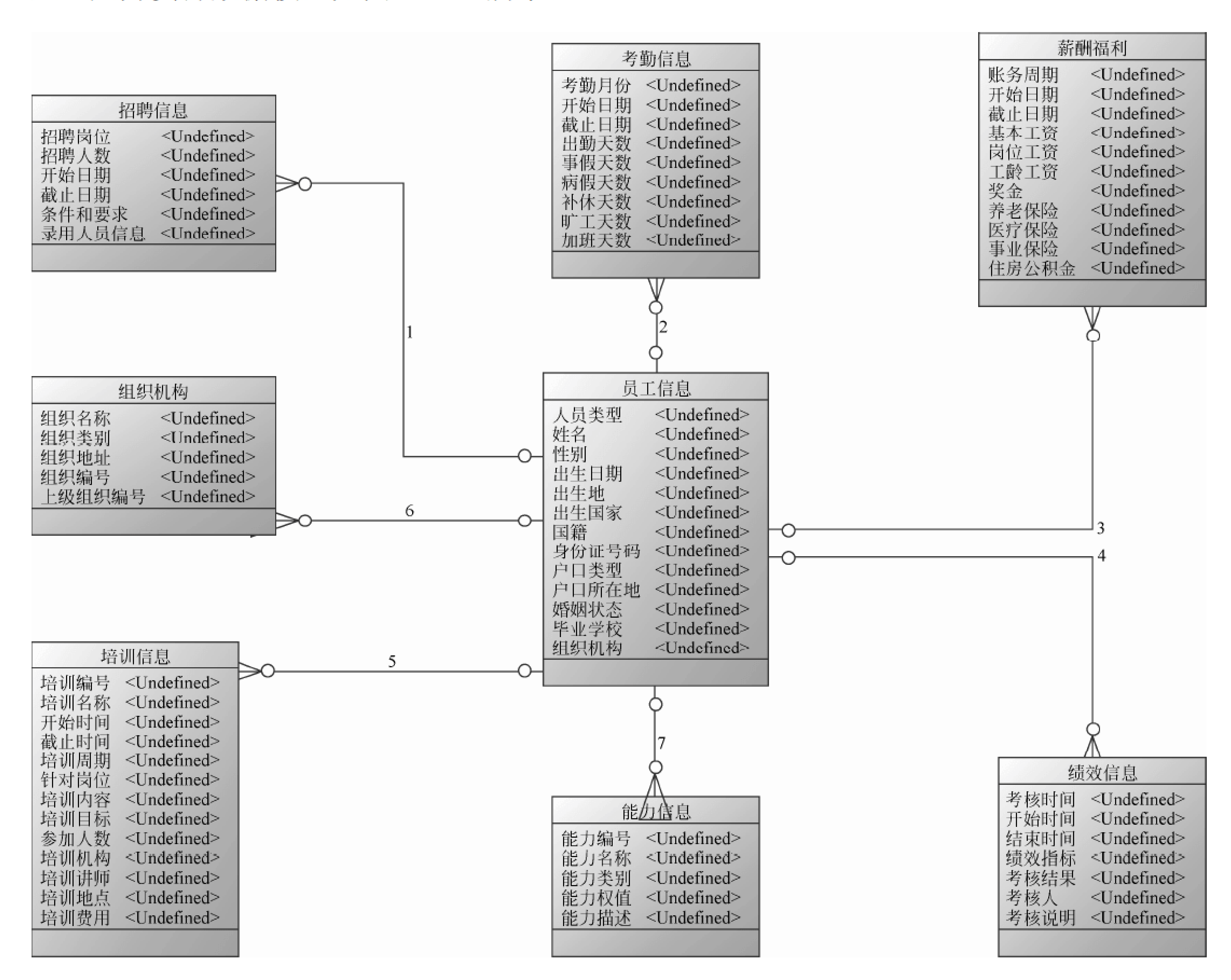


图 3-4-11　企业人力资源数据模型示例

在互联网发达的今天，企业有条件从互联网获取关于所需人才的更多有价值的数据，企业应当借助大数据，构建、丰富和完善人才库，打造具有竞争力的人才队伍。

8. 财务数据模型设计

财务数据是计算企业收入和成本的基础，为了便于记账，企业借助会计科目定义财务

发生的对象，形成财务计算的基础。

企业财务数据模型如图 3-4-12 所示。

在移动互联网时代，社会专业化分工更细，企业需要在发挥自身核心竞争力的同时，与更多的原材料供应商、服务提供商、分销商等合作，为了降低企业风险，企业管理者需要及时掌握成本收益情况。可见，企业非常有必要构建面向移动互联网的财务模型。

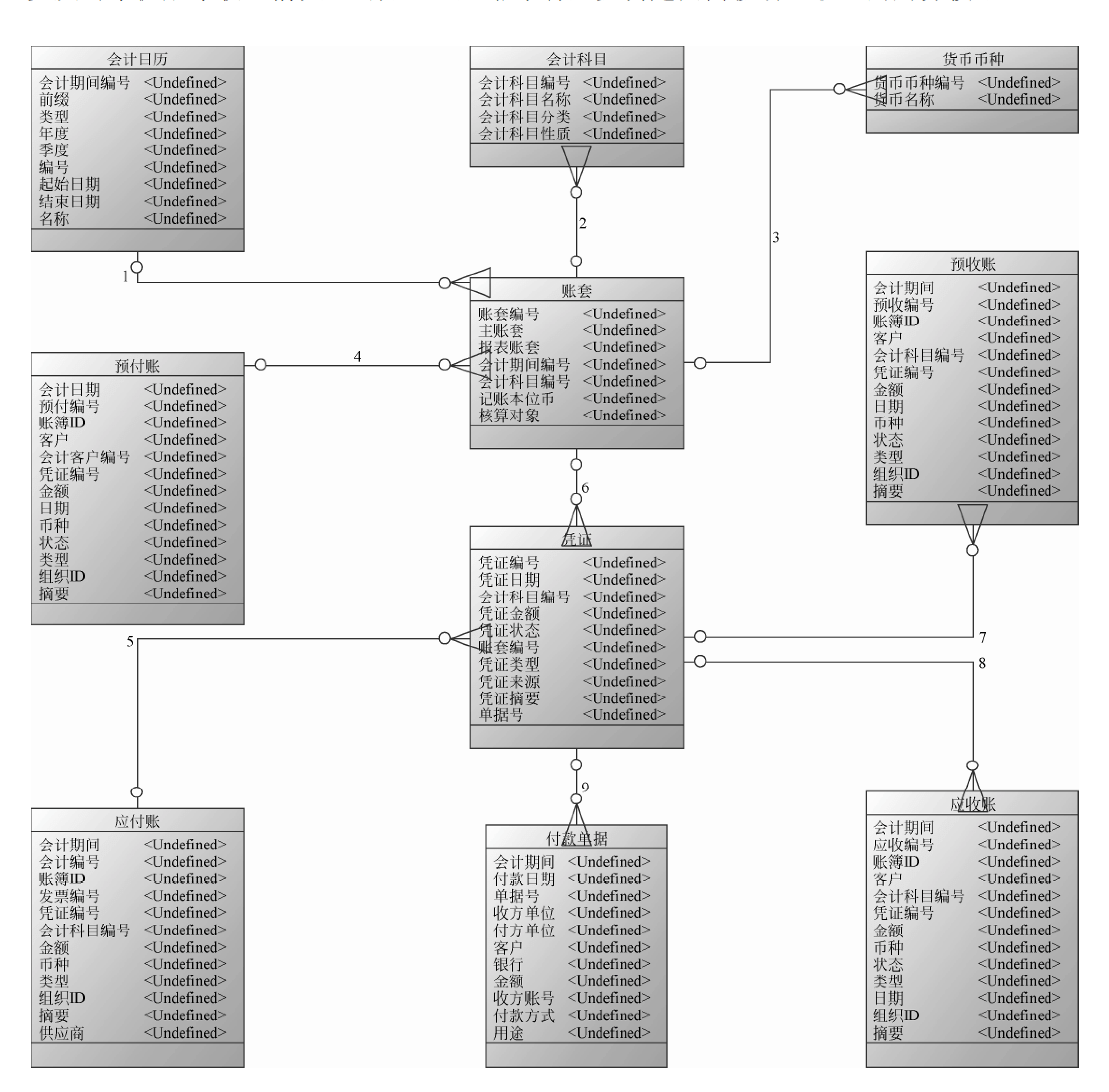


图 3-4-12　企业财务概念模型示例

9. 资产数据模型设计

资产与资源的关注点不同。资产关注物的价值属性，而资源则关注“物”的使用属性。

在企业的生产和运营过程中，资产会随着时间的推移、技术的革新等逐渐贬值。因此，企业通过全生命周期管理资产，可以准确地评估成本，降低企业运营风险。

企业资产数据模型如图 3-4-13 所示。

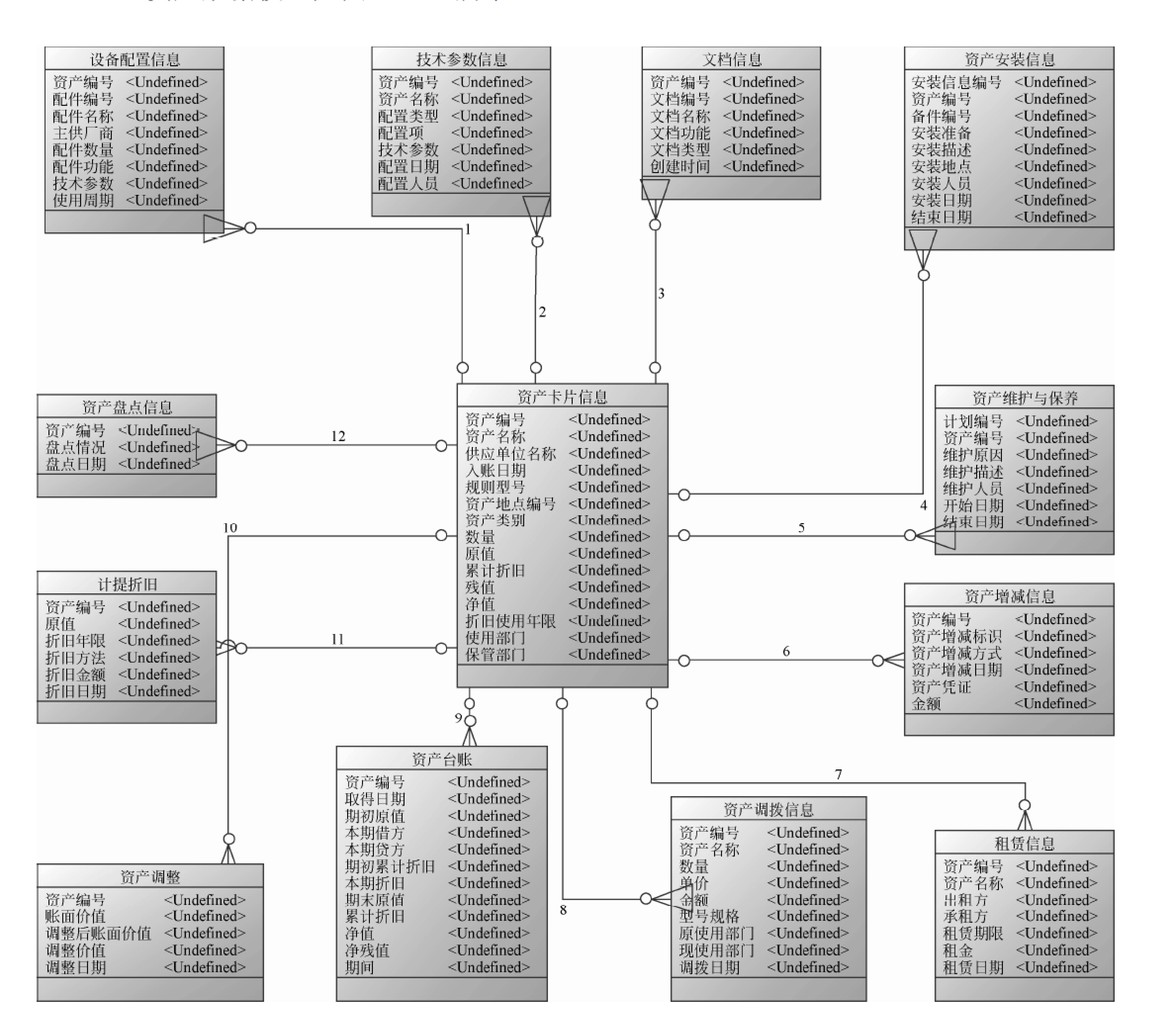


图 3-4-13　企业资产概念模型示例

3.4.2 面向分析的数据模型设计

面向操作的数据模型用于支撑企业高效地完成各种业务活动，而面向分析的数据模型则是为了帮助企业发现数据背后隐藏的规律。

著名数据仓库专家 Bill Inmon 和 Ralph Kimball，从不同角度提出了分析模型设计方法。Bill Inmon 主张基于细颗粒度的原始数据构建分析模型，这样分析人员可以具有更大的自主性。Ralph Kimball 则从用户需求出发，主张构建面向主题的多维模型，使得分析模型更加贴近用户，更容易使用。

两位数据仓库大师从不同的视角提出了分析模型的构建思维与方法，一个是为数据分析人员提供更大的灵活性，一个是为了更好地满足数据分析人员的个性化需求，两者各有利弊。下面简单分析一下面向分析的数据，即多维模型构建的方法。

1. 多维模型建模方法

数据仓库大师是多维数据模型设计的倡导者。Ralph Kimball 将多维模型设计步骤归结为四步法，如图 3-4-14 所示。

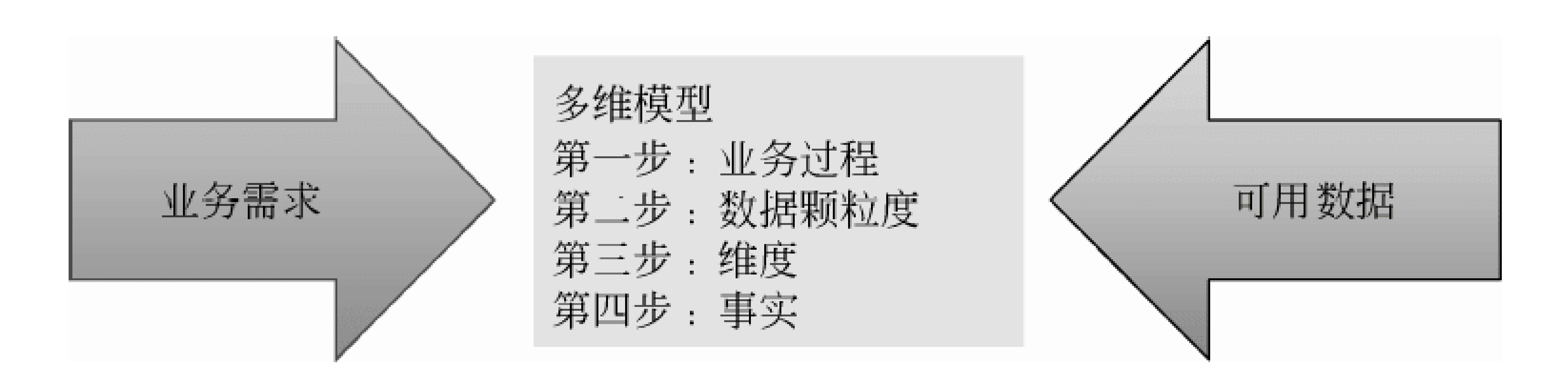


图 3-4-14　多维模型设计四步法

从图 3-4-14 可以看出，多维模型设计的前提是业务需求和可用数据，这与 3.2 节中大数据服务的需求分析方法是类似的，业务需求决定了多维模型的目标和方向，而可用数据则是多维模型设计的前提和基础。

多维模型设计四步法如下。

第一步：选择业务过程。企业以面向过程管理的方法论为指导，通过业务过程来刻画企业生产经营的各种业务活动。采用面向过程而不是面向职能的管理方法，消除了职能部门之间的流程难以贯通、信息不能充分共享等问题。

从时间轴看，业务过程分为企业战略管理、基础设施生命周期管理、产品生命周期管

理、运营准备、服务开通、服务保障以及计费收费几个阶段。通过对企业业务过程的分解，可以清晰地看到业务过程在整个业务过程框架体系中的位置。例如，选择产品定价和商品促销两个业务过程，通过构建多维模型满足企业决策分析的需要。

第二步：定义数据颗粒度。数据颗粒度表明了事实表中每一行代表的含义。例如：

- 每行代表客户在超市购物的每一笔交易；
- 每行代表每个员工每天的考勤记录；
- 每行代表每个移动用户每天的上网流量；
- 每行代表每个人每月的公交卡刷卡次数；等等。

预先声明数据的颗粒度是非常重要的，事实表中数据的颗粒度表示了多维模型中装载数据的规模以及数据分析的能力。通常是数据颗粒度越小，越能够从更多维度对数据进行分析，而分析的性能与较大颗粒度的数据相比也会差一些。

第三步：识别维度。用户可以通过多个维度来查看数据的统计特征，正如“横看成岭侧成峰，远近高低各不同”。通常企业对数据统计的维度包括日期、产品、渠道、促销、雇员、支付方式等。

第四步：识别事实。事实是数据分析的“结果”，比如销售数量，销售额度等。比如，企业统计 2015 年第三季度通过各种销售渠道销售的 iPhone 6 终端数，那么统计的维度就是日期维度（2015 年第三季度）、渠道维度（实体营业厅、网上营业厅等），而事实就是销售的 iPhone 6 终端数。

通常采用 SQL 对多维模型进行分析，SQL 实现代码如下：

```
-- 1.构建日期维度测试数据
DROP TABLE dates PURGE;
CREATE TABLE dates(date_key INTEGER,date_str VARCHAR2(8),quar_num VARCHAR2(8));
INSERT INTO dates(date_key,date_str,quar_num) VALUES(1,'20150401','第三季度');
INSERT INTO dates(date_key,date_str,quar_num) VALUES(2,'20150402','第三季度');
INSERT INTO dates(date_key,date_str,quar_num) VALUES(3,'20150403','第三季度');
COMMIT;
SELECT * FROM dates;
-- 2.构建销售渠道维度测试数据
CREATE TABLE channel(chan_key INTEGER,chan_type INTEGER,chan_desc
VARCHAR2(20));
INSERT INTO channel(chan_key,chan_type,chan_desc) VALUES(1,1,'实体营业厅');
INSERT INTO channel(chan_key,chan_type,chan_desc) VALUES(2,2,'网上营业厅');
INSERT INTO channel(chan_key,chan_type,chan_desc) VALUES(3,3,'电话营业厅');
```

```
COMMIT;
SELECT * FROM channel;
-- 3.构建销售事实表测试数据
DROP TABLE sales facts PURGE;
CREATE TABLE sales facts(seq no INTEGER,sale date VARCHAR2(8),chan_type
VARCHAR2(20),product VARCHAR2(20),quantity INTEGER);
INSERT  INTO  sales facts(seq no,sale_date,chan_type,product,quantity)
VALUES(1,'20150401',1,'iPhone6',60);
INSERT  INTO  sales facts(seq no,sale_date,chan_type,product,quantity)
VALUES(2,'20150401',1,'iPhone6',50);
INSERT  INTO  sales facts(seq no,sale_date,chan_type,product,quantity)
VALUES(3,'20150402',2,'iPhone6',40);
INSERT  INTO  sales facts(seq no,sale_date,chan_type,product,quantity)
VALUES(4,'20150402',2,'iPhone6',30);
INSERT  INTO  sales facts(seq no,sale_date,chan_type,product,quantity)
VALUES(5,'20150403',3,'iPhone6',20);
INSERT  INTO  sales facts(seq no,sale_date,chan_type,product,quantity)
VALUES(6,'20150403',3,'iPhone6',10);
COMMIT;
SELECT * FROM sales facts;
-- 4.执行基于多维模型的事实统计
SELECT A.QUAR NUM,B.CHAN DESC,SUM(C.quantity) PROD_QUAN_SUM FROM DATES
A,CHANNEL B,SALES FACTS C
WHERE A.DATE STR = C.SALE DATE AND B.CHAN_TYPE = C.CHAN_TYPE
GROUP BY A.QUAR_NUM,B.CHAN_DESC;
-- 5.统计结果
```

QUAR NUM	CHAN DESC	PROD QUAN SUM
第三季度	电话营业厅	30
第三季度	实体营业厅	110
第三季度	网上营业厅	70

上面的示例中，GROUP BY 后的字段就是各个统计维度，而 SUM 内的字段就是事实表中需要统计的“事实”。

2. 多维模型的两种结构

面向分析的数据模型包括星型和雪花型两种结构。

星型结构以事实表（Fact Table）为中心，外围是各种维度表（Dimension Table）。事实表的主要特点是包含数字数据（事实），并且这些数字信息可以汇总。每个事实表包含一个由多个部分组成的索引，该索引是事实表的外键，是相关维度表的主键。星型结构的多

维数据模型如图 3-4-15 所示。

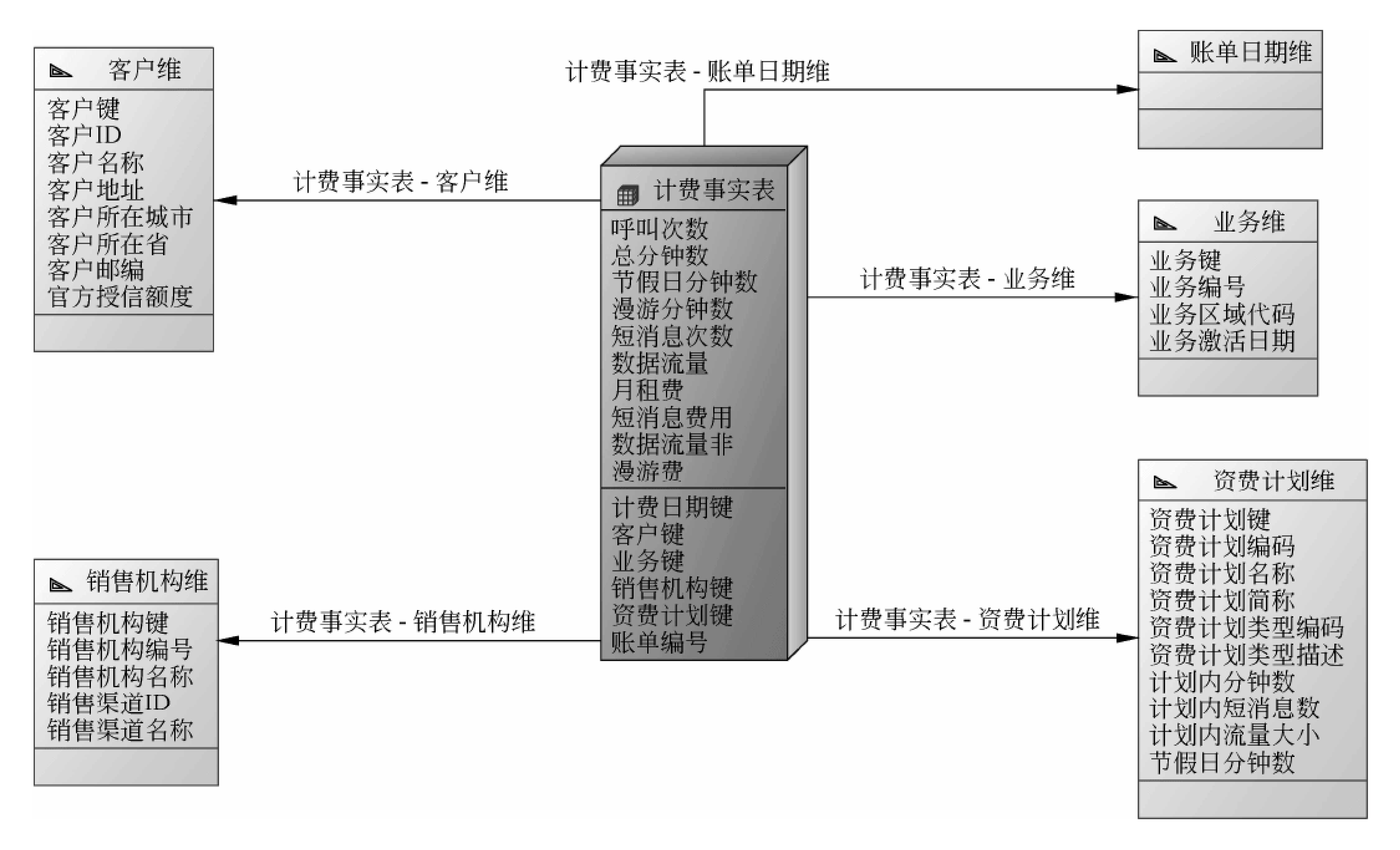


图 3-4-15　星型结构的多维模型示例

从图 3-4-15 可以看出，星型结构的多维模型是以计费事实表为中心的，包括账单日期、客户、业务、资费计划、销售机构几个维度。在多维模型设计时，尽量减少维度表的数量，根据经验，维度表尽量不要超过 20 个，如果维度表太多，应当将几个维度表合并为一个较大的维度表，太多的维度表会因为多表之间复杂的关联关系而大大降低统计的整体性能。

雪花型结构是星型结构的组合，不同的事实表由一个或者多个公共的维度表连接起来。

3.4.3　大数据服务元数据设计

什么是元数据？元数据（Meta Data）是数据的数据，既包括数据结构定义，也包括数据操作过程。元数据对于数据分析人员而言，就像一本大词典，要想写出一篇好的文章，必须学会词语的含义，将多个词语连接在一起，就变成了一篇好文章。

面向操作的数据库中存放的元数据包括数据表结构、主键、外键、索引、分区等，这些元数据通过数据定义语言（DDL）来构建，借助数据操作语言（DML）实现对数据内容的存取。面向操作的数据库用于支撑事务型应用，主要是开发者关注元数据的定义，最终

用户无须关注元数据的定义，只需使用事务处理系统提供的功能即可。

与面向操作的数据库不同，面向分析的数据仓库用于实现决策型应用，数据分析人员需要掌握元数据，才能够对不同的数据进行关联，发现数据背后隐藏的规律。著名数据仓库专家 Bill Inmon 将数据仓库分为交互区、集成区、近线区和归档区。在数据管理策略的指导下，数据在不同区之间移动。数据在移动的过程中，需要将 ETL 规则作为元数据存储，以便理解数据转化前后语义的变化。

3.5 大数据服务容量设计：海纳百川，有容乃大

与事务处理应用相比，大数据服务属于分析处理应用，由于两者的数据处理特点不同，因此容量估算方法也有一定的区别。

随着时间的推移，会有越来越多的数据进入数据仓库，如果不及时管理存储空间，大数据服务就会难以运行。

为了完成大数据服务的容量设计，需要进行容量的估计、容量占用监测以及容量调整。企业可以根据大数据的规模、分析时长要求等估计大数据服务所需的存储空间、计算能力以及网络传输带宽。

在大数据服务运行的过程中，要根据监测到的容量占用情况，及时迁移或删除数据、增加基础设施资源等，以保障大数据服务的正常运行。可以根据数据活跃度、存储时限规则等将数据转移到相应的存储设备。

1. 事务处理系统容量设计方法

事务就是请求提交到返回结果的过程。面向操作的应用的特点为事务性。事务的 4 个特性为 ACID，即原子性、一致性、隔离性、持久性。比如在网上购物，填写完了商品、配送信息并完成支付后，单击提交就发起了一个交易申请，然后系统会给出交易结果。要求事务处理系统能够快速地响应请求，通常是几秒钟之内，否则系统用户是无法接受的。

基于数据仓库构建的在线分析处理（OLAP）与面向操作的事务型应用相似，为了解决 OLAP 应用快速响应用户的问题，通常采用构建中间表的方式，预先将分析结果放入中间表，然后系统从中间表中直接取出分析结果。

面向操作的事务处理应用需要估算计算、存储和传输三个方面的能力，能力估算方法如下：

1）事务处理应用计算能力估算方法

计算能力需求 = 计划支撑的用户数×单用户的交易量×单个用户需要的 TpmC(tpm 是 transactions per minute 的简称，C 指 TPC 中的 C 基准程序)×冗余系数。

对于 HP、IBM 等服务器设备厂商，通常会给出某个配置下其服务器的 TpmC 能力，因此可以根据估算结果和厂家某个型号配置的服务器 TpmC 能力的对比，算出需要某种品牌型号服务器的数量。

2）事务处理应用存储能力估算方法

存储能力需求 = 计划支撑的用户数×单用户产生的记录数/天×单条记录大小×冗余系数，此外存储空间估算还应当考虑操作系统、中间件、索引、日志等额外占用的空间以及 RAID、数据存储时间策略等因素，最后再根据磁盘类型、容量来选择所需的硬盘数量，磁盘分为 SATA 盘、SAS 盘、FC 光纤盘等，磁盘容量通常包括 300GB、450GB、1TB 等。

3）事务处理应用网络能力估算方法

网络能力需求 = 计划支撑的用户数×单用户传输带宽×冗余系数。批量数据传输往往需要较大的网络带宽。可以根据带宽要求，选择光口还是电口，采用千兆端口还是万兆端口。

从事务处理应用的容量估算方法可以看出，计划支撑的用户数是对面向操作的事务处理应用进行容量设计时考虑的主要因素。

2．大数据分析处理系统容量设计方法

与事务处理应用相比，大数据服务属于分析处理应用，由于两者的数据处理特点不同，因此容量估算方法也有一定的区别。大数据服务通常要经过数据 ETL、数据存储、数据分析、数据展示、数据开放的过程，因此在计算能力、存储能力以及网络能力的估算上也有自身的特点。大数据服务在不同阶段对于基础设施的需求如图 3-5-1 所示。

从图 3-5-1 可以看出，对于一个普通的大数据项目，通常要经过数据采集（1）、数据存储和数据转换（2.1，2.2，3.1，3.2，3.3，3.4）、数据展示（4.1，4.2）三大步骤，具体处理过程如下。

第一步：从各种数据源采集数据。数据源分为内部和外部数据源两种。内部数据源是企业自身的数据，比如电信运营商的用户上网数据是从交换机获取的业务使用记录；外部数据源是企业从外部获取的数据，比如移动终端配置数据是从第三方公司数据库获取的。

采集数据的方式也分为主动和被动两种。主动方式是主动去数据源抓取数据，比如可以通过网络爬虫在各大网站获取数据；被动方式是企业为数据源设定好存储位置，让数据提供方按照时间策略向指定位置存放数据。

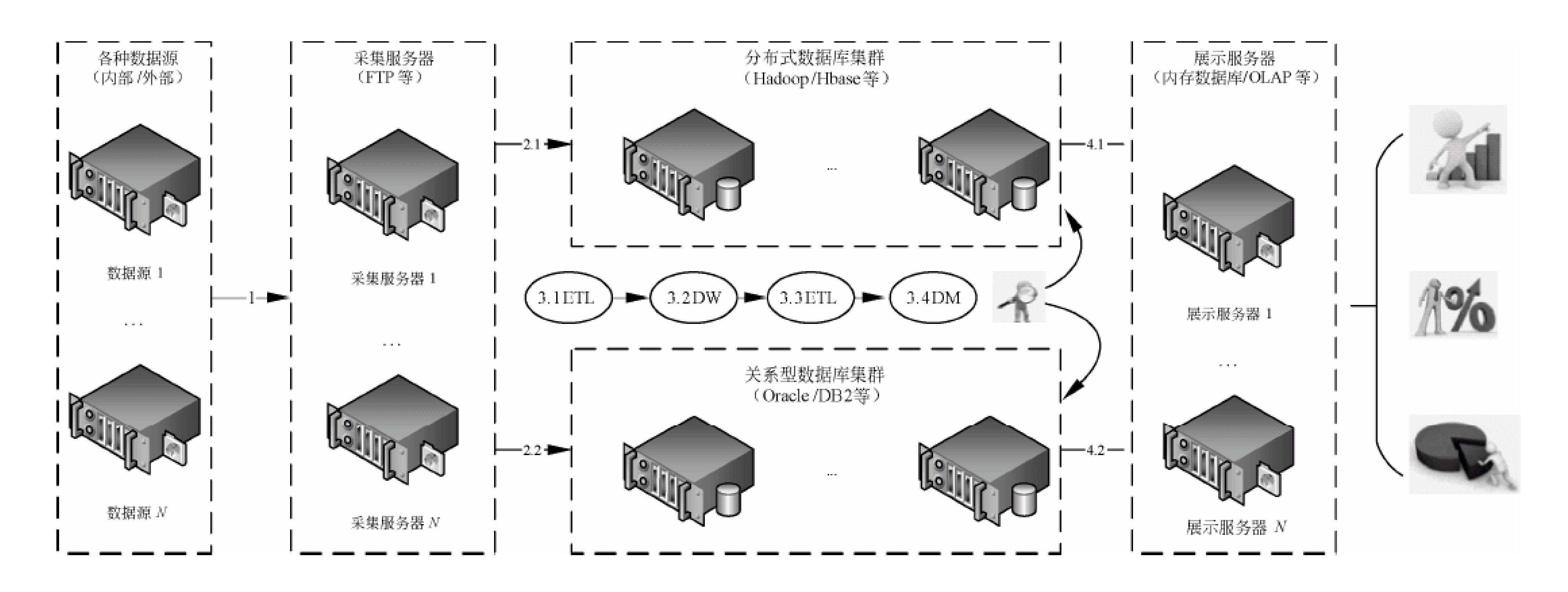


图 3-5-1　大数据服务不同阶段的基础设施需求

第二步：数据存储和数据转换。企业可以根据数据特点不同采取不同的数据存储策略，如果数据规模大或者预期的数据规模大，传统的关系型数据库是无法满足快速处理要求的，因而需要考虑采用分布式数据库，比如 Hadoop/HBase。类似 Hadoop/HBase 这样的分布式数据库的特点是扩展性好，如果存储空间不够，只需增加存储服务器即可。不足之处是 HBase 只适合单表或者多表之间关联关系简单的场景，对于需要数据操作或者多表关联的应用，还是需要基于关系型数据库实现的。

关系型数据的优势就是能够对数据进行整合和统计，从而使得用户可以从多个维度来查看分析结果。当然，由于关系型数据库基于单机模式完成架构设计，尽管也可以支持集群方式部署，但是横向扩展能力有限。可见，多表关联查询要比键值映射方式对数据库管理系统的要求高，但是没有键值映射方式的扩展性好。因此，在大数据存储时，需要结合应用需求和数据库存储特征来进行综合考量：使用分布式数据来存储数据规模大、增量大并且以数据查询为主的数据，采用关系型数据库实现需要多表关联的查询统计功能。

当原始数据存储到数据库中以后，需要对数据进行抽取、转换与加载，保证数据质量和应用要求。数据过程通常是经过初步的 ETL，然后将数据存储到数据仓库，接着再次对数据进行 ETL，将数据加工成面向不同主题的数据集市，以便于从多个维度查看数据统计结果。

第三步：数据展示阶段。虽然已经费了很大力气完成了数据的抽取、转换、丰富等工作，但是数据毕竟是给人看的，数据展示得越好，越容易让用户看到数据背后隐藏的事实和规律。比如电信运营商为了查看各地区数据流量的多少，可以基于电子地图，不同数据流量区间用不同的颜色标识，这样可以直观地看到各省数据流量的多寡。

1）大数据分析处理系统容量估算方法

大数据分析处理系统容量估算可以分为理论估算法和实验估算法两种类型。

理论估算法的数据基础包括文件数、单个文件数的记录条数、单条记录大小、数据采集周期，数据采集周期包括一次、一天、一个月等，这样就能够算出某个时间段内的总数据量大小。然后再考虑磁盘的冗余空间系数，就可以算出对于磁盘空间总的需求量。理论估算法适合于没有样本数据的场景。

理论估算法的计算公式为：存储空间大小 = 文件个数×单个文件记录数×单条记录大小×时间长度×冗余系数。

实验估算法基于某个时间段的样本数据。用户可以用操作系统自带的命令查看文件大小。如果进入数据仓库的数据从时间上是连续的，则可以通过样本数据测量值与时间长度相乘，算出大数据分析处理系统存储空间需求。

实验估算法的计算公式为：大数据分析处理系统存储空间大小 = 样本数据量大小×时间长度×冗余系数。

2）大数据分析处理系统计算能力估算方法

传统数据处理与存储架构是“主机+磁盘阵列”的集群方式，主机可以是小机、PC 服务器或者刀片服务器，磁盘阵列可以是 NAS、SAN 等，采用的协议可以是 FC、IP 等。

传统数据处理与存储架构解决了存储资源和计算资源的共享问题。多个服务器组成的集群可以将计算资源统一管理，接收请求的负载均衡器会根据服务器负荷将请求发送到计算资源充足的服务器。磁盘阵列实现共享的方式更加容易理解，就是多个磁盘放到一个机箱中，机箱可以扩展并且机箱内可以热插拔磁盘，这样可以便于扩展磁盘空间。

“主机+磁盘阵列”的系统架构将计算和存储分离，通过计算群和存储群的方式提高了并行处理能力，满足了高并发的事务处理应用的系统要求，但是这种架构也带来了新的问题，就是计算和存储资源的横向扩展能力是有限的。

大数据服务的特点是数据量大，尤其是随着时间的推移，数据量会不断增大，要求计算和存储资源能够具备几乎没有限制的扩展能力。为了满足不断增加的数据量，谷歌公司提出了基于 MapReduce 和 GFS 的分布式计算架构，与“主机+磁盘阵列”的架构方式不同，

谷歌公司利用廉价的机器设备，通过软件将能力不一的大量计算机设备连接到一起，降低了 IT 基础设施采购成本，提升了 IT 基础设施的扩展能力。随后，Apache 受谷歌的 GFS/MapReduce 架构的启发，提出了 Hadoop 分布式计算架构。

可见，新型的面向大数据的分布式计算架构与“主机+磁盘阵列”的系统架构在设计思路上完全不同，大数据计算能力估算的方法也是不同的。

3.6　大数据服务过程设计：卓有成效的管理者

大数据服务过程包括服务目录管理、容量管理、可用性管理、连续性管理、服务等级管理、信息安全管理、供应商管理等。

在设计方法方面，大数据服务与支撑企业运营的服务既存在区别，又存在联系。不同之处是：大数据服务的设计主要以“数据”为参考点，“数据”类型越多、越丰富、越新鲜，越有助于设计好的服务；两者的共同点是：大数据服务归根结底还是为企业运营服务的，是为了提升企业在建设、市场营销、产品销售、客户服务、企业管理等方面的能力。

大数据服务在设计阶段的过程包括服务目录管理、容量管理、可用性管理、信息安全管理、供应商管理等。

3.6.1　大数据服务目录管理

服务目录相当于饭店里点菜的菜单，用户通过服务目录可以看到有哪些服务，管理者也可以通过服务查看服务所依赖的资源，进而可以算出服务的成本效益。

随着大数据服务数量的增多，需要对其进行分级分类管理，以便能够快速地检索和定位大数据服务。同样，大数据服务也会不断优化完善，因此需要对大数据服务增加版本标签的方式予以区分。

大数据服务目录可以按照大数据服务支撑的业务应用进行分类组织，比如一级大数据服务可以分为投资建设类、市场经营类、资源运营类、行政综合类和企业管理类。可以在一级基础上进一步细分，比如市场经营类可以细分为市场营销、销售、客户服务和计费收费。按照这种分类方式，可以明确大数据服务支撑的业务应用所在的位置，可以让使用者

更高效地找到大数据服务。比如，某大数据服务的目标就是支持企业的网络规划设计，那么就应当在投资建设类中查找满足这一要求的大数据服务。

3.6.2 大数据服务容量管理

容量是组织的 IT 资源提供服务能力的吞吐量。IT 资源所提供的容量衡量指标包括支持的最大并发用户数，最大在线用户数，服务器最大计算能力，最大存储空间，最大网络出口带宽等。

容量管理不仅对于 IT 服务设计重要，对于大数据服务设计仍然非常重要。

大数据服务通常以三种形式提供，不同类型的大数据服务对于容量的需求是不同的。

第一类是支持操作决策的大数据服务，这样的大数据服务需要嵌入生产型过程中，用户在使用生产型应用的过程中同样会调用大数据服务，因此要求大数据服务能够提供保证业务连续性的能力。这种类型的大数据服务与面向操作的事务型应用对于容量的需求类似，因此对该类大数据服务进行容量设计时，可以把其当作事务型服务看待。

第二类大数据服务属于统计分析型，该类大数据服务更多是为了满足企业中层管理人员统计某个时间段的数据，比如统计某个季度的产品销售数据、某个年度的现金流量等，辅助管理人员发现生产经营中存在的问题，由于这样的大数据服务并不嵌入生产型应用之中，因此对于实时性要求没有那么高，对该类大数据服务进行容量设计时，主要考虑特定时间段的容量需求，比如月初月末。

第三类大数据服务主要面向企业的高层战略管理人员，比如企业的总经理、战略规划师等，这样的人员通常关注半年以上的中长期规划，需要借助大数据服务掌握市场情况、与竞争对手之间的差距等，这样的大数据服务对于响应时间通常要求不高，更关注数据背后隐藏的规律，设计重点为决策模型，由于该类大数据服务往往需要以多年的历史数据为分析基础，因此可以考虑采用基于云架构的基础设施，以便弹性地适应不断增长的支撑能力需求。

3.6.3 大数据服务供应商管理

在大数据时代，数据成为企业的核心资产，而由于社会的专业化分工，数据也势必分散在不同的企业之中。大数据服务与企业其他原材料一样，如果不能保证及时准确地提供，

将会降低大数据服务的能力。可见，对于大数据服务的供应商进行有效管理同样非常重要。

企业引入大数据服务与企业引入生产型服务一样，需要进行有效的管理。比如大数据服务供应商的准入和退出管理、服务质量管理、服务绩效管理等。供应商管理的目的是企业能够及时、有效地获取到满足要求的大数据服务，包括数据提供的时效性、数据质量等。

供应商的准入管理主要对供应商大数据服务提供能力的要求，降低大数据服务提供风险，企业可以通过与供应商签署大数据服务供应合同，从法律上保证因供应商不能按照要求提供服务带来的损失，降低企业生产和经营风险。以定期对供应商提供的大数据服务进行考核评价作为大数据服务是否退出的依据。

3.6.4　大数据服务安全管理

将大数据看作企业核心资产的同时，也就意味着数据在企业中有着非同一般的价值和作用。此外，大数据还有不同于企业其他资产的独特性，比如企业的客户数据会涉及个人或者企业的隐私，可能会涉及企业的商业秘密。

为了保证大数据服务的安全性，需要从以下三个方面做起。第一是保证数据不会被非法获取，企业可以通过权限控制机制实现认证和授权。第二是当企业或个人使用数据时，要进行数据使用记录，保留“痕迹”，为审计工作做好准备。第三是数据的对外提供采用匿名或者统计数据的方式，保证数据使用方不会看到真实的个体数据，如果确实需要则可以采用审批和合约的方式，在法律制度上对数据予以保护，要严惩违法者。

3.6.5　大数据服务等级管理

服务等级是大数据服务的用户和大数据服务的提供者之间的共同约定，大数据服务的提供者需要按照约定的服务等级来提供服务。

当大数据服务提供方并没有按照约定的服务等级提供服务时，需要进行服务能力提升，以保证按照约定的服务等级提供服务。比如，服务等级中约定用户从提交大数据服务请求到服务响应的时间为 3 秒钟以内，如果用户实际使用过程中没有满足这样的服务等级要求，大数据服务提供方则需要确认信息系统的容量设计是否存在问题，如果存在问题，则可以通过扩容大数据服务基础设施容量的方式来满足性能要求。当然，在大数据服务提供方没有为用户提供相应等级的服务时，大数据服务提供方应当给予使用方一定的经济

补偿。

服务等级除了在系统响应性能方面的要求外，主要还是大数据服务提供的数据质量是否能够满足要求。例如，数据完整性和数据准确性应当保证超过合同约定的百分比。应当预先建立双方都能认可的数据质量验证方法。

3.6.6 大数据服务可用性管理

服务的可用性直接关系到用户的体验。如果用户体验好，则会提高用户的办事效率，反之则有可能导致用户的流失并减少企业收入，可见可用性管理是非常重要的。

大数据服务分为三类：嵌入生产过程中的服务、提供决策参考的服务以及提供趋势预测的服务，以上三类大数据服务对于可用性的要求是有差别的。

对于嵌入生产过程中的大数据服务，需要保证高可用性，否则会因为无法及时做出决策而影响企业的生产经营，比如某银行的贷款业务流程中集成了信用评估服务，而信用评估服务就是一个大数据服务，只有当信用评估大数据服务输出客户的风险敞口后，才能确定是否能够为客户提供贷款以及贷款额度，如果信用评估大数据服务不可用，则会延长用户获得贷款的时间，从而降低银行贷款业务的办理效率，甚至导致客户的流失。

比较而言，提供决策参考的大数据服务和提供趋势预测的大数据服务，对于响应的实时性要求相对较低，因此对大数据服务的可用性要求相对也较低。当然，它们对于可用性的要求也需要根据具体情况来判断，如果企业应付突发和紧急情况，这时候对以上两类大数据服务的可用性要求也是非常高的，如果大数据服务不可用，则会为企业带来很大的损失，原因是大数据服务的不可用影响到组织的决策效率，错失了调整经营策略的好时机。

可见，大数据服务可用性对于企业的生产经营都是非常重要的，需要通过可用性管理的方法和手段来保证大数据服务具有较高的可用性。

实现大数据服务高可用性的方法分为两种类型：被动型和主动型。

被动型方法是要求系统对于大数据服务的运行情况进行实时监控，根据监控结果进行量度和分析，并通过报表形式展现分析结果，根据分析结果来定位和解决影响大数据服务可用性的故障点。

主动型方法是采集用户使用和系统运行数据进行主动分析，预测可能影响大数据服务可用性相关的问题，提前优化和完善，防患于未然。

3.6.7 大数据服务连续性管理

顾名思义，服务连续性管理就是保证服务不间断。对于面向操作的事务型应用，服务的连续性是服务质量的重要考量指标，当服务出现故障后应当尽快发现和解决问题，服务恢复时间的长短体现了服务连续性管理水平的高低。

对于大数据服务而言，与生产流程结合紧密的大数据服务的连续性是需要重点考虑的类型。对于其他类型的大数据服务，应该重点保障数据采集服务的连续性，因为如果数据采集失败就意味着大数据服务依赖的数据样本减少，进而影响到数据分析的结果。

3.7 大数据服务组织设计：分工不分家

按照专业化分工和关注点分离的原则，大数据服务业务分析师和大数据服务系统架构师是两个非常重要的角色。

在大数据服务设计阶段，需要的角色主要包括大数据服务业务分析师和大数据服务系统架构师。

大数据服务业务分析师的职责是关注大数据服务如何满足业务需要，如何提升企业在战略管理、建设管理以及运营管理方面的能力。

大数据服务系统架构师的职责是关注大数据服务如何落地，采用哪种架构方式，需要多少基础设施资源等。

3.7.1 大数据服务业务分析师

业务分析师负责大数据服务的发现、定义以及业务测试。大数据业务分析师基于可以获取的大数据资源，结合企业过程框架，来发现什么样的大数据服务可以支持企业更好地完成战略、建设以及运营工作。大数据业务分析师的职责包括：

（1）发现可能为企业构建大数据服务的数据源；

（2）基于各种数据源，结合企业过程框架，发现和定义大数据服务；

（3）对大数据服务进行测试，验证其是否能够支持企业战略、建设和运营。

3.7.2 大数据服务系统架构师

大数据服务系统架构师负责基于先进适用技术，完成对大数据业务分析师定义的大数据服务的设计方案。

在大数据服务创意到大数据服务实现之间，大数据服务系统架构师起到桥梁和纽带的关键作用。大数据服务系统架构师需要对于大数据技术非常了解，同时对于大数据服务也有深刻的理解。大数据服务系统架构师的职责包括：

（1）对大数据服务进行架构设计，包括技术架构、功能架构、集成架构等方面的设计；

（2）根据大数据服务需求，结合各种大数据相关技术，对大数据服务进行原型设计和实现；

（3）跟踪大数据相关的各种技术，大规模海量数据要求大数据技术能够满足数据处理的高效性，同时也要求借助适用的大数据技术，从多种类型的数据之中发现更大的价值。

3.8 主要内容回顾

“孕育”意味着埋下一颗希望的种子，要想让这个“种子”满足预期要求，必须从全局和长远考虑，对于设计大数据服务这颗高科技“种子”，还应当具备正确的思维方式，具备面向服务、面向过程、全生命周期、数据即资产的观念，将大数据作为服务或者产品来对待，以价值创造为衡量大数据服务的原则，整合企业内部及社会数据，充分挖掘大数据的潜力。

不同于面向操作的事务型应用，大数据服务更多的是一个探索发现的过程，对于已经发现规律并模型化的大数据服务，可以构建数据模型并嵌入事务型应用的过程环节中，比如银行对于个人客户的授信服务，可以构建授信模型，通过收集个人客户相关数据计算客户风险敞口，实现客户贷款过程中的快速授信。

在很多情况下，大数据服务是一个探索发现的过程，即通过不断地尝试，发现数据之间的规律。因此对于大数据服务，应当采取快速迭代、螺旋上升的开发模式，通过不断调整和优化数据模型和算法，达到大数据掘金的目的。

大数据服务是先有“数”后有“求”的，因此应当首先对大数据的潜在能力进行分析。比如当具备移动用户上网记录大数据后，参考移动用户上网记录大数据的元数据，发现移动用户上网记录中包含用户数据（终端、号码、IP 等）、应用数据（域名、IP 等）以及网络数据（位置区、小区、网络类型、流量、时长、经纬度等），从而确定借助通信大数据，可以具备“再现”移动用户的上网行为的能力。

对于“求”，可以分析一下可能有哪些需求。还是以移动用户上网记录大数据为例，网络规划设计是对网络建设进行决策，那么需求就是如何完成无线网络的规划设计，比如在哪里建设，在哪里需要扩容，建设或者扩容的规模有多大等。有了移动用户上网记录大数据，就可以基于用户价值和应用价值完成无线网络的规划设计了。

大数据服务架构设计是通过制定大数据服务参考框架，理清大数据服务在不同阶段、不同层次上的关注点以及这些关注点之间的关系。

大数据服务模型设计关注面向操作和面向主题的数据模型设计，通过数据模型的构建，解决大数据承载、数据分析以及数据展现问题。

如果说数据是大数据服务构建的基础，那么数据模型则是大数据服务实现的载体，数据模型的设计对于大数据服务至关重要。通过分析大数据服务从操作型数据模型到分析型数据模型的渐进过程，通过清晰地看到数据模型从操作环境到分析环境发展的变化，加深对数据模型的认识。

大数据服务容量设计则是关注如何规划和监控大数据服务基础设施资源需求，以最佳成本效益的方式完成大数据基础设施能力的设计。

数据通常要经过采集、存储、整合、分析、展示、归档、销毁的过程。从大数据的价值角度看，那些活性高，频繁使用的数据，通常具有较长的生命周期，反之，那些很少被使用的数据，尽管由于法律法规要求需要保留较长的时间，但是其应当“离休”，迁移到“非活动”区域。

从成本角度看，应当综合大数据活性、价值、法律法规要求等对数据进行分级存储，实现成本效益的最大化。比如，经常使用的数据放在价格高但是访问速度快的缓存、内存、磁盘中，而将那些偶尔访问的数据放在廉价的磁盘、光盘等存储介质中，对于访问频率极低甚至很长一段时间（比如 3 年）没有访问的数据，应转移到价格更低的磁带介质中。

通过对大数据活性、价值、法律法规的观察，将数据存储到不同的存储纸质上，即提高了数据访问的效率，也降低了大数据存储成本。对于那些已经确定不用的或者按照存储要求到期的数据，通过审批机制进行数据销毁。

大数据服务过程设计的目的是保证大数据服务能够得到有效的管理。大数据服务过程设计主要包括大数据服务目录管理、容量管理、供应商管理以及安全管理四个方面。

目录管理过程保证大数据服务能够得到最大程度的共享和使用，消除企业内部大数据服务能力交叉和重叠的现象，企业在形成大数据服务之前，需要查看是否已经具备类似的大数据服务，尽量重用现有的大数据服务。

容量管理过程保证大数据服务拥有足够的、最佳成本效益的基础设施资源，包括存储空间、计算能力以及网络传输带宽。企业可以基于大数据访问活跃度、法律法规要求等完成数据的迁移、归档、销毁等任务。

供应商管理过程用于保证数据源的质量和及时性。大数据时代，组织势必会引入多个供应商的数据，供应商提供数据的质量和及时性关乎组织的数据分析能力，企业应当建立对供应商数据质量的评价方法和制度，保证数据的准确性。供应商提供数据的及时性对组织大数据服务生产效率影响很大，同样是需要保障的重要因素。

安全管理过程保证大数据的合规性，组织的大数据往往是个人和组织在生产生活中留下的“痕迹”，因此组织对隐私侵犯和商业秘密侵犯的分析和管理，成为大数据服务“开放”或者“封闭”的重要依据，组织可以通过匿名、审批、统计数据提供等方式规避隐私侵犯和商业秘密侵犯问题。

大数据服务组织设计在大数据服务的构建过程中起到了非常重要的作用。由于大数据服务通常是在探索中发现的，因此要求大数据服务设计人员具有关于问题域很强的专业背景甚至多行业知识背景。

按照专业化分工和关注点分离的原则，大数据服务设计阶段主要考虑两种角色：大数据服务业务分析师和大数据服务系统架构师。

大数据服务业务分析师主要关注专业领域分析模型的构建，这个角色要求具有很强的行业知识，能够根据组织决策需求进行建模，能够使用数据建模和数据分析软件和工具完成分析模型的构建、优化、展现工作。

大数据服务系统架构师关注大数据服务基础设施的架构，包括存储架构、计算架构以及网络架构，保证大数据服务具有可用的基础设施资源，负责监控数据的活动情况，根据大数据服务需求进行数据迁移或者增加大数据服务基础设施资源，满足大数据分析的性能需求和存储空间需求。

第 4 章
分娩：从幕后到台前的华丽转身

“十月怀胎，一朝分娩”，大数据服务经过漫长的“孕育”阶段，在母体中逐步发育，成为一个待产的“婴儿”，但是这个“婴儿”要从母体降临到人间，还需要经历一个“分娩”过程。

对于大数据服务这个“待产婴儿“而言，能否符合用户要求，能否实现从幕后到台前的华丽转身，还需要经过一个痛苦的、充满期待的过程。

大数据服务要完成从开发测试环境生产环境的转换，需要经历集成测试、系统测试、用户接受测试、系统部署、编制文档、用户培训这几个阶段。

集成测试、系统测试、用户接受测试的前提是大数据服务从开发环境部署到测试环境。集成测试主要完成大数据各个部分之间的接口测试，比如 ETL 系统、分析系统、展示系统之间的集成测试。系统测试是从整体上对大数据服务进行测试，分为功能性测试和非功能性测试，比如系统设计阶段的系统功能点是否符合设计要求，系统性能、可靠性、可用性、扩展性、安全性方面是否满足设计要求等。测试阶段通常是从数据仓库中提取一部分数据作为测试数据，因此测试阶段通常仅仅是在一定程度上对大数据服务进行验证。尽量采用脚本完成系统的模拟测试，减少因为人为随机输入引起的错误。

用户接受测试主要是找出一部分大数据服务的用户参与测试，因为这些用户是了解业务的，会从业务视角出发审视大数据服务，因此会比开发和测试人员更能够发现大数据服务存在的问题。通过让用户在测试阶段试用大数据服务，可以提前发现大数据服务中存在的问题，降低项目风险。

当完成集成测试、系统测试、用户接受测试以后，需要将大数据服务从测试环境迁移到生产环境。在此阶段，需要编制文档和用户培训工作。

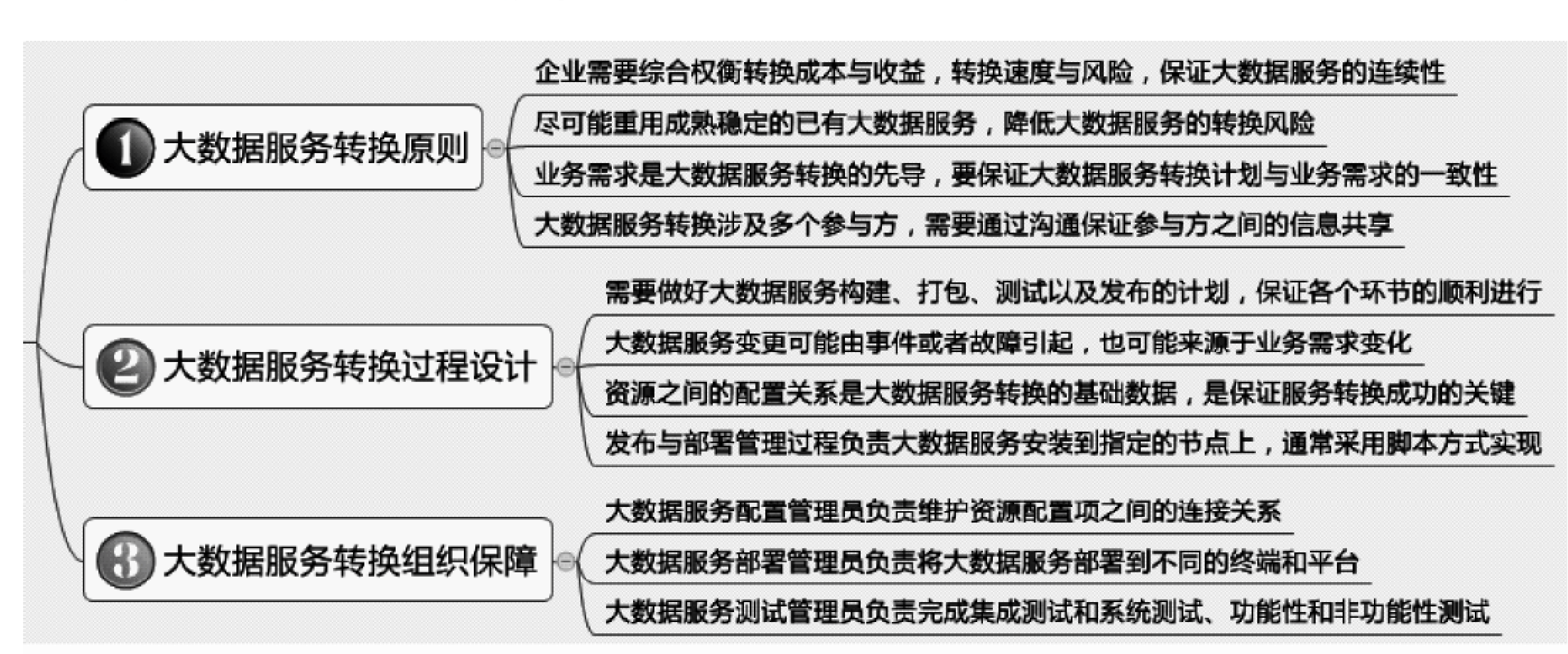
编制文档一方面可以备查，另一方面也便于各参与方沟通交流。文档可以分为核心和外围两种类型。核心文档主要为元数据定义文档，包括表结构、存储过程、视图、触发器、功能等的定义。外围文档包括大数据服务使用说明书、用户培训 PPT 等。

用户培训的目的是教会用户如何使用大数据服务的功能。培训用户的人员最好是负责大数据服务展示的人员，因为负责大数据服务展示的人员更了解大数据服务的功能如何使用。大数据服务功能展示设计的特点之一就是要便于用户从多个维度观察，发现数据背后隐藏的规律，因此培训材料要让用户掌握从多个维度来查看数据分析结果的方法。

此外，系统安全管理也是大数据服务转换阶段需要重点测试的，要保证不同的大数据服务能够按照预先设定的数据权限使用数据。在应用层面，要完成系统的账户管理、认证管理、授权管理和审计管理，完成用户、组织、岗位、角色、权限的影射和维护。

需要转换的大数据服务包括两种类型：新的大数据服务和变更的大数据服务，不同类型的大数据服务在转换方法和过程上是不同的。

本章内容思维导图如下所示：



4.1 大数据服务转换原则

大数据服务转换充满了期待又存在着风险和挑战，需要综合权衡转换成本与收益、转换速度与风险。

大数据服务转换既充满了期待又存在着风险和挑战，如果没有正确的转换策略和原则作为指导，那么大数据服务转换很可能会失败。

大数据服务转换需要综合权衡转换成本与收益、转换速度与风险。对于新的或者变更的需求，如果要使转换成果带来的成本大于收益，则应当重新考虑是否进行大数据服务转换。同样，如果大数据服务没有充足的转换时间，则会因为没有进行充分的分析和设计而产生大数据服务不可用的风险。

为了保证大数据服务能够成功转换，需要有几个关键原则作为指导：最大复用原则、服务转换计划与业务需求保持一致原则、与干系人保持良好沟通原则。

1. 最大复用原则

大数据服务转换会存在风险，如果服务转换失败会给企业带来损失，因此大数据服务转换尽可能使用已有的过程和系统，开发大数据服务复用规范以及引入行业最佳实践，以提高服务转换的成功率。

最大复用原则既可以保证大数据服务转换的效率，又可以通过引入经过实践检验的、高质量的转换过程而提高转换的成功率，企业尽可能借鉴企业或外部第三方已有的大数据服务转换经验，包括软件代码、转换脚本等。

2. 服务转换计划与业务需求保持一致原则

业务需求是用户价值的体现，因此需要将服务转换计划与业务需求保持一致，否则服务转换是没有价值的。

与业务需求保持一致原则强调业务需求作为大数据服务转换的重要前提条件。无论开发面向操作的事务型应用还是面向决策的分析型应用，业务需求始终是系统努力的方向和

目标，越早地发现业务需求存在的问题，就能够越早地发现和降低大数据服务转换的风险。

3．与干系人保持良好沟通原则

与大数据服务干系人保持良好的沟通同样非常重要。大数据服务必须满足干系人的需要才能体现其价值，如果客户、用户等干系人没有很好地理解新增或者变更的大数据服务对他造成的变化，会降低大数据服务转换的成功率，因此需要与干系人及时地沟通，还要保障干系人能够及时获取到所需的相关文档，如大数据服务操作手册等。

大数据服务是多个干系人共同参与完成的，包括业务用户、大数据服务分析师、大数据服务架构师、大数据服务部署工程师、大数据服务测试工程师等，如果不同角色的人员获取的信息不对称，就难以保障大数据服务能够成功转换。

采用行业最佳实践计划并管理在打包、部署、测试以及发布阶段所需的资源，保证按照预期的成本、质量以及时间将大数据服务成功转换到正常运营状态。

此外，企业还需要对客户、用户等干系人进行培训，以便其能够更好地使用大数据服务。在大数据服务未正式运营之前，监控并量度大数据服务的使用效果并与预期效果进行对比，及时发现大数据服务存在的问题并进行改进完善。

4.2 大数据服务转换过程

大数据服务转换过程包括转换计划、变更管理、资产与配置管理、发布与部署管理、验证与测试、评估以及知识管理。

“变化是永恒的，唯一不变的是变化”，随着用户对于大数据服务认识的深入，会不断提出新的需求，而新的需求要求大数据服务重新设计、开发并转换生产运营状态。

大数据服务转换的驱动力分为三种：外部业务需求引起、技术发展变化引起和适用企业管理新要求引起。

大数据服务转换过程的目标是实现大数据服务成功地转换到预期的状态，为此需要配置管理和知识管理作为支撑。

配置管理可以对大数据服务的连接关系进行管理，以便进行大数据服务的部署实施，

并帮助发现大数据服务部署过程中引起错误的故障点。

知识管理可以帮助企业积累大数据服务管理过程中的经验，形成知识库，更加快速高效地解决大数据服务管理过程中遇到的问题。

大数据服务转换过程主要包括转换规划与支持、变更管理、服务资产和配置管理、发布和部署管理、服务验证与测试、评估、知识管理。

4.2.1 大数据服务转换计划

在大数据服务转换之前，需要做好大数据服务构建、打包、测试以及发布的计划，使得新的或者变更的服务能够顺利地投入生产。

大数据服务转换计划过程包括：

（1）对服务转换进度、变化、问题、风险以及偏移进行管理；

（2）与客户、用户等干系人沟通、改进并完善服务转换绩效。

4.2.2 大数据服务变更管理

在大数据服务运营过程中，通过事件、故障、问题等确定需要变更的大数据服务。变更的需求也可能来自于业务需求的变化，业务需求的变化会导致大数据服务设计的变化，从而形成大数据服务变更的需求。

企业可以根据变更的紧迫性分为不同的优先级：立即、高、中、低。变更的优先级通常由大数据服务对于企业效益和风险的影响程度确定，对于企业价值高的大数据服务具有较高的变更优先级。

4.2.3 大数据服务资产与配置管理

企业要完成大数据服务转换，需要以服务配置关系为基础。例如，某个软件服务部署在哪个中间件上，中间件部署在哪个操作系统上，操作系统部署在哪个主机上，主机位于哪个网络中，网络如何接入通信网络中。大数据服务的分层部署结构如图 4-2-1 所示。

从图 4-2-1 可以看出，大数据服务通常由 5 层来承载。简化起见，图中没有单独标识

位于第三层内部的虚拟化层和位于第四层内部的支撑平台层。

与面向操作的应用一样，大数据服务需要操作系统、主机设备、存储设备、网络设备作为底层的基础设施，大数据服务与面向操作的应用的不同之处是大数据服务需要以满足海量数据存储要求的分布式计算与存储架构为基础，比如采用 MapReduce 作为大数据计算架构，Hadoop/HBase 作为大数据存储架构。

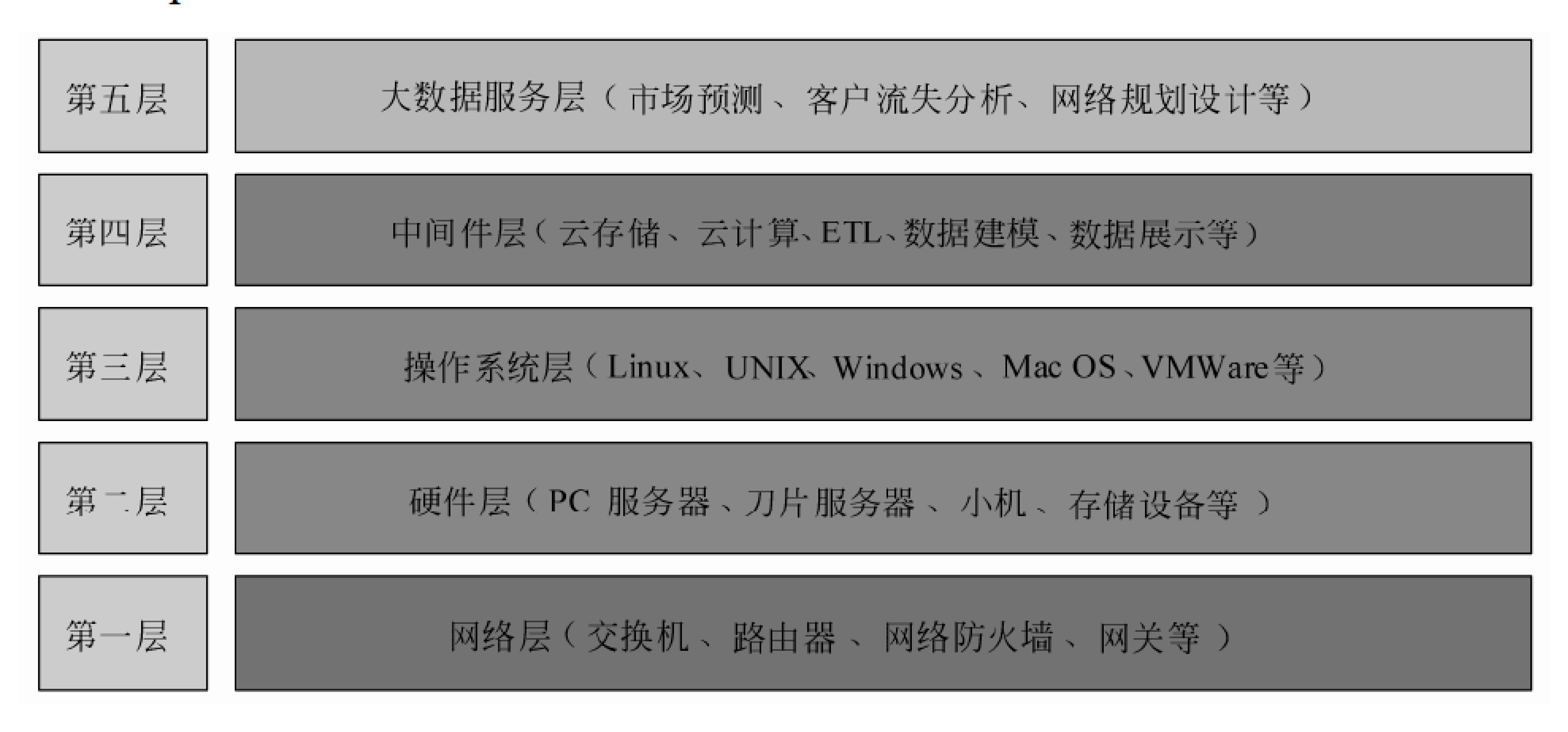


图 4-2-1　大数据服务部署层次结构图

下面以“客户流失分析”大数据服务为例，分析该大数据服务在各个层次的部署要求。

第一层（网络层）：与事务型应用相比，大数据服务具有大量的数据传输特性，因此对于网络带宽的要求比较高，建议构建独立的网络基础设施，与事务型应用分开，以免降低企业事务型应用的响应速度，影响客户体验。

第二层（硬件层）：与事务型应用相比，大数据服务对于操作系统没有特别的要求，因为大数据服务在操作系统的上层实现资源的分配、调度和管理。

第三层（操作系统层）：事务型应用的特点是数据操作频繁，但是单次数据传输量较小，因此通常采用“主机+磁盘阵列”的集群架构，主机集群可以满足大量用户高并发的需求，当用户访问操作型应用时，应用会通过负载均衡器，将请求发送到主机集群中负荷低的计算节点，而磁盘阵列则可以保证数据的可靠性。本质上是通过冗余空间换取可靠性，磁盘阵列包括 RAID0+1、RAID5、RAID6 等多种级别，在主机集群和磁盘阵列之间通常采用光纤通道（Fiber Channel,FC）的方式进行连接，光纤通道具有高达 1Gb/s 级别的传输速率，能够满足主机和磁盘阵列之间的数据传输要求。

为了应对海量数据存储的要求，大数据服务采用了与事务型应用不同的计算和存储架

构。事务型应用采用“计算”和“存储”分开集群的方式，而大数据服务则采用“计算”和“存储”一体化集群的方式。前者通过在存储区域网络中增加磁盘的方式提升基础设施能力，属于纵向扩展，系统整体性能不会随着主机以及磁盘阵列的增多而线性提升，扩展能力有限，因而难以满足快速增长的海量数据存储要求。

而“计算”+“存储”一体化的集群架构则没有以上限制，原理上是一体化集群架构借助分布式存储和分布式计算软件实现基础设施资源的调度和管理。这种架构采用一体化集群的主机节点作为集群新的能力，在一体化内部建立容错机制，以保证数据的可靠性。

第四层（中间件层）：大数据服务在这一层与事务型应用的实现方法和过程方面有很大的区别。

事务型应用面向操作，采用“主机+磁盘阵列”的集群架构，通常是在主机集群中部署应用中间件和数据库中间库，如果系统采用 B/S 结构，则部署 Web 中间件。处理过程为：用户进入应用系统并发起请求（比如录入数据后提交订单）到主机集群，主机集群内部的负载均衡器将请求转发到服务器并执行计算操作，应用服务器再与数据库服务交互，最终通过数据库服务器将数据存入磁盘阵列。

大数据服务的处理过程为：首先，从各种数据源采集数据，经过 ETL 进入大数据集群服务器。然后，经过 ETL，将数据存入数据仓库。接着，经过 ETL 将数据导入数据集市，最后，通过数据展示中间件展示到终端设备上。

对于简单的查询应用，可以通过部署在大数据集群服务器上的云计算和云存储中间件获取所需数据。对于统计报表应用和复杂的数据分析应用，需要对大数据进行 ETL 并将数据加载到传统的关系型数据仓库，最后再借助数据展示中间件在终端上展现出来。

第五层（大数据服务层）：应用层功能是基于第四层中间件实现的，比如市场预测分析、客户流失分析、网络规划设计等。

4.2.4 大数据服务发布与部署管理

随着大数据服务的不断完善，会形成多个面向不同场景的版本，为了便于大数据服务的发布和部署，需要对大数据服务进行版本管理，同时制订部署计划，保证即使在部署失败的情况下仍旧能够为客户提供不间断的服务。

大数据服务与事务型应用在部署结构上是不一样的，两种类型的应用的部署结构对比如图 4-2-2 所示。

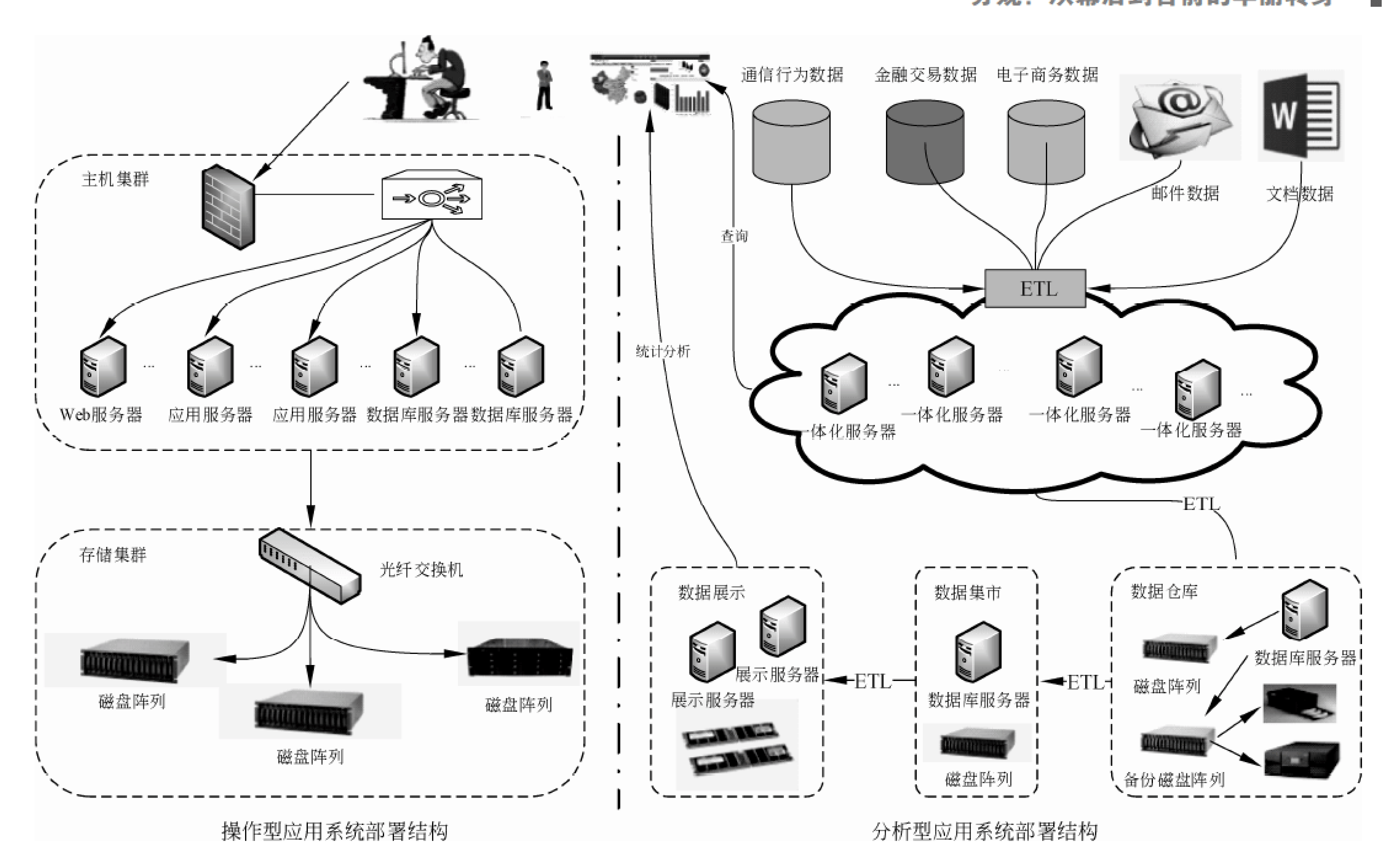


图 4-2-2　大数据服务与事务型应用部署结构对比

从图 4-2-2 可以看出，操作型应用采用“主机+磁盘阵列”的集群方式，这种系统部署结构适用于单次操作数据量小但操作频繁的事务型应用，通过构建主机集群，基本可以满足亿级用户的访问请求，同时由于事务型应用产生的数据规模一般为 TB 级，通过扩容磁盘阵列即可满足存储空间需求。此外，由于事务型应用产生的数据为企业的核心生产经营数据，对于数据可靠性要求高，因此适用于 RAID 方式来保证数据的可靠性。

事务型应用的部署过程为：首先，将主机服务器、磁盘阵列、网络设备等安装设计方案部署到机房的机架内，通过光纤、网线等连接起来并加电。其次，在硬件设备上安装操作系统，通常操作系统和设备管理软件已经由设备提供商在硬件设备预装，如果不符合要求，则需要按照设计要求安装所需操作系统和设备管理软件。再次，在操作系统之上安装各种中间件，比如负载均衡代理软件、Web 中间件、交易中间件、数据库管理软件、管理软件等。最后，将预先打包部署的应用软件部署到应用中间件上，在数据库管理软件上创建表、视图、存储过程、触发器、函数等并导入测试数据。

当以上过程完成后，就可以执行集成测试、系统测试、系统使用文档编制、用户培训、用户接受测试等测试工作了。

对于大数据服务，由于数据量可能会在 PB 级以上，数据量太大，“主机+磁盘阵列”的系统部署结构无法通过横向扩展满足计算和存储需求，因此需要采用“计算和存储一体化”的系统部署架构，通过分布式存储管理软件和分布式计算软件实现一体化服务器集群。如果数据存储空间不足，可以通过增加一体化服务器实现计算能力和存储能力的提升。

尽管一体化服务器集群能够满足大规模数据存储和计算需求，提供高效的数据查询服务，但是复杂的统计分析还需要借助传统的关系型数据仓库实现。基于关系型数据库构建数据仓库的基本过程为：首先，将企业内部和外部的数据经 ETL 装载到数据仓库。然后，根据不同部门、不同角色的需要，形成面向不同主题的数据库，比如面向市场营销部门的市场主题库，面向财务部门的财务主题库等。最后，为了提高数据展示的速度，需要部署展示服务器和基于内存的报表分析软件，在终端上展示分析结果。

大数据服务在第一步和第二步的部署方法和内容与事务型应用的部署过程基本一致，都是完成硬件的安装上架、操作系统软件的安装以及各个硬件平台之间的网络联通。

大数据服务第三步部署与事务型应用差别很大。主要差别是大数据服务要部署大数据存储、计算、ETL、展示等专用中间件。例如，在一体化服务器上要部署分布式存储和分布式计算中间件，比如 Hadoop/HBase、MapReduce 等，在执行数据 ETL 的服务器上部署 ETL 工具。部署数据仓库管理软件、数据集市管理软件、在展示服务器上部署基于内存的报表展示中间件，等等。

大数据服务部署的第四步是在各个中间件上安装和配置各种应用软件，包括基于 Hadoop 的数据查询应用软件、ETL 脚本和应用软件、基于数据仓库和数据集市的存储过程、基于数据展示中间件的界面展示应用软件等。

4.2.5 大数据服务验证与测试

当系统硬件、操作系统、管理软件、中间件、应用软件、测试数据等部署完成后，意味着已经完成了测试环境的搭建，具备了执行测试的条件。

测试阶段分为集成测试、系统测试、用户接受测试三个阶段。

集成测试主要测试设备之间、部署在不同设备上的应用软件之间是否可以互通，是否符合预先设定的接口定义。比如，当网络设备、主机、存储设备等上架并连接好网线后，系统集成商要测试网络设备之间是否能够相互通信；中间件产品提供商要测试部署在不同

操作系统之上的中间件是否可以正常通信，例如 Web 中间件和数据库管理软件之间是否可以正常连接；应用软件提供商要测试部署在不同设备上的应用软件之间是否可以正常发送和接收消息，不同应用软件之间是否按照预先定义的接口要求传递消息，等等。

系统测试是从整体上对系统功能、性能、可靠性、安全性、可伸缩性等方面进行测试的。系统测试的方法包括边界测试、正常范围测试、各种异常情况测试等。以系统功能测试为例，可采用输入符合要求的数据、不符合要求的数据的方法测试系统功能是否满足要求；对于系统性能的测试，可以通过模拟多用户场景，实施压力测试，查看系统是否能够承载某个规模用户的并发请求，是否满足系统响应时间要求；对于系统可靠性测试，可以通过移走集群中部分主机的方式测试集群是否满足可靠性要求，可以按照地域范围，测试系统是否满足同机房内部、同城内或者异地之间的容灾需求；对于安全性，可以采用网络攻击模拟软件测试系统是否发现异常网络行为，并采取断开连接、加入黑名单等方式阻止异常网络行为。对于应用级安全测试，可以测试账户是否可以被非法窃取，是否可以通过审计功能来发现异常操作，对非正常使用系统功能和数据的行为进行预警。

用户接受性测试是让最终用户参与测试，通常经过三个月的用户接受性测试后，系统进入正常运行状态。

4.2.6 大数据服务评估

在大数据服务正式运营之前，需要对大数据服务进行评估，判断其是否是可以接受的、是否具有价值等，减少大数据服务在运营阶段带来的风险和损失。

4.2.7 大数据服务知识管理

知识是人类根据以往的经验，对发现的问题以及解决问题的方法进行总结而形成的。例如，通过上学获得了知识，而这些知识是前人在对于生产与生活中的客观世界的认识中形成的。

对于大数据服务而言，知识管理的作用是为了帮助企业将信息在恰当的时间地点传递给有此需要的人，以便快速解决遇到的问题。知识管理是大数据服务各个阶段都需要的，每个阶段都需要借助知识管理过程来解决问题。

企业可以构建大数据服务知识库管理平台，将生产经营中形成的或者外部学习的经验教训知识化。知识管理平台对于研究、咨询、设计等知识型企业尤为重要，这类企业在多年的生产经营过程中形成了丰富的经验，如果将这些知识经验进行有效管理，就可以方便员工查询、使用。一方面可以提高工作效率，保证工作质量，另一方面也可以增强员工归属感，降低员工流失率。

4.3 大数据服务转换组织设计

大数据服务转换中涉及的角色主要包括资产管理员、配置管理员、配置分析师、部署管理员、测试管理员。他们默默无闻，却担负着将梦想变为现实的重任。

大数据服务转换过程需要多个角色共同参与完成，这些角色可以由一个或者多个人员来担当，人员配比可以根据企业大数据服务所需的工作负荷而定。

大数据服务转换的过程中可以存在不同的角色。比如在配置管理过程中，可以设置服务资产管理员、配置管理员和配置分析师角色，在发布和部署过程中，可以设置部署管理员角色，在服务测试与验证过程，可以设置测试管理员角色。下面就简单分析一下不同角色的职责。

4.3.1 大数据服务资产管理员

大数据作为企业的核心资产，需要进行有效的管理以便企业能够对大数据服务产生的效益以及消耗的成本进行计算。通过成本效益分析可以确定该大数据服务是否可行，需要占用企业多少成本等。

大数据服务由多个“资产”组成，这些“资产”可以是有形的基础设施（硬件），也可以是看不见摸不着的无形资产（软件）。需要借助一些方法和手段来衡量大数据服务资产的价值。比如，对于网络、服务器、存储设备等硬件设备，可以通过资产原值和折旧的方式来计算资产现值。像大数据服务中的软件资产，更多是由人的智力创造的，需要通过人员数量、人员单价、软件算法复杂系数等来估算。

大数据服务资产管理员的主要职责包括：

（1）定期评估大数据服务资产的价值，要求掌握软件价值评估的方法和工具；

（2）将大数据服务资产评估结果上报给企业管理人员，对于可能会带来企业经营风险的资产，需要及时上报给企业管理人员。

4.3.2 大数据服务配置管理员

在有些情况下，企业需要新增或者变更大数据服务，这时需要对新增或者变更的大数据服务进行转换，而大数据服务转换需要掌握大数据服务资产资源的配置。为了保证大数据服务能够成功地转换到运行状态，需要对大数据服务进行配置管理。

大数据服务是由应用、平台、基础设施等不同层面的资源支撑实现的，因此，大数据服务配置管理的任务就是维护它们之间的真实关系。当新增大数据服务时，需要维护大数据服务相关的软硬件资源与现有软硬件资源的配置关系。如果是大数据服务变更，则需要更新大数据服务影响的资源的配置关系。

大数据服务配置管理员的主要职责包括：

（1）负责维护大数据服务相关的资源、能力、价值等信息；

（2）负责维护大数据服务相关资源之间的连接关系。

4.3.3 大数据服务配置分析师

新增或者变更大数据服务往往会影响很多配置项（Configuration Item，CI），而数据维护不及时或者数据录入错误会导致不正确的配置关系，这些错误的配置关系会对以后新增或变更大数据服务造成麻烦，因此需要大数据服务配置分析师来及时发现配置关系中存在的问题并进行纠正，以保证顺利地完成大数据服务的新增或者变更。

大数据服务配置分析师的职责主要包括：

（1）定期对大数据服务配置项以及它们之间的关系进行评估，纠正其中存在的问题；

（2）提供以月度、季度、年度为单位的配置分析报告，说明配置失败的原因以及解决的方法，并将配置管理经验作为知识库的一个重要输入，同时将配置分析报告上报给企业管理者，作为绩效考核的依据。

企业信息系统内部资源之间的配置关系是否准确，对于企业生产经营的效率具有非常

重大的影响，因此建议企业在资源管理方面予以重视，通过规范化的管理流程予以保障，并将资源管理纳入绩效考核。企业可以定期进行资源资产关系稽核，及时发现资源配置关系中存在的偏差并进行纠正，为大数据服务提供准确的配置关系数据。

4.3.4 大数据服务部署管理员

当按照设计要求完成大数据服务的开发与单元测试后，就可以对开发成果进行打包、构建，然后发布到测试环境和生产环境了。大数据服务部署管理员需要准备大数据服务所需的基础设施资源，并将大数据服务部署到相应的资源节点。

大数据服务部署的前提是资源之间正确的配置关系，首先应当安装大数据服务所需的硬件基础设施，然后安装大数据服务所需的系统软件，包括操作系统、中间件、数据库等，最后将大数据服务应用软件部署到相应的系统软件之上。

大数据服务部署管理员的职责主要包括：

（1）完成大数据服务的构建与打包。在不同的平台上，大数据服务依赖的资源是不一样的，应当首先准备好与大数据服务兼容的软硬件资源，使其可以构建为一个可运行的应用；

（2）完成大数据服务的版本管理。当大数据服务完成 Bug 的修改或者新需求的变更后，会相应地升级大数据服务的版本。如果大数据服务的版本在运行中出现问题，还可以回退到正常的版本，保证大数据服务的连续性。

4.3.5 大数据服务测试管理员

当构建并部署大数据服务后，就可以对大数据服务进行测试了。大数据服务测试的目的是验证其是否可靠、可用，安全性、可伸缩性、性能等方面是否符合企业要求。

大数据服务测试管理员的主要职责包括：

（1）对大数据服务进行功能性测试，验证是否满足设计阶段的功能要求，功能性测试的方法通常是根据测试案例，验证在预设输入的情况下是否产生期望的结果；

（2）对大数据服务进行非功能性测试，非功能性需求包括便捷性、可用性、扩展性、安全性、性能等方面，与功能性测试相比，非功能性测试的难度更大，需要模拟各种失败

的场景。

4.4 主要内容回顾

为了保证大数据服务从开发测试环境顺利转换到生产环境，需要遵循一些简单的原则。首先，最大化复用原则，保证大数据服务能够快速、稳定地交付。其次，要确保大数据服务转换计划与业务需求的一致性。最后，要通过沟通、培训等方式，保证大数据服务转换的参与方能够高效地协同配合。此外，企业需要综合权衡转换成本与收益，转换速度与风险，保证大数据服务的连续性。

当完成大数据服务的设计和开发工作之后，还需要经过打包、测试、部署等一系列转换过程，才能实现大数据服务的上线运行。

大数据服务转换过程包括转换计划过程、变更管理过程、资产与配置管理过程、发布与部署管理过程、验证与测试过程、评估过程以及知识管理过程。

转换计划过程需要完成大数据服务的构建、打包、测试以及发布计划的制订，保证各个环节的顺利进行。

变更管理过程由大数据服务运行过程中产生的事件或者故障激发，也可能由新的需求驱动。

资产与配置管理过程是大数据服务能否转换成功的前提。当大数据服务部署任务完成后或者资源变更后，应当及时更新配置项之间的连接关系。

发布与部署管理过程负责大数据服务安装到指定的节点上，通常采用 Ant、Maven 等脚本方式实现大数据服务的自动化构建、打包与部署。

大数据服务转换阶段需要多个不同分工的角色。大数据服务配置管理员负责维护资源配置项之间的连接关系。部署管理员负责将大数据服务部署到不同的终端和平台。测试管理员负责完成大数据服务的集成测试和系统测试、功能性测试和非功能性测试。

第 5 章

培育：调整、巩固、充实、提高

当小家经过“筑巢”、“联姻”、“孕育”、“分娩”的过程后，“大数据服务”终于降落到人间，一家人看着可爱的小宝宝，自然特别高兴。当然，父母含辛茹苦，将小宝宝养大成人是非常不容易的，何况每个父母都望子成龙，望女成凤，就更加不容易了。

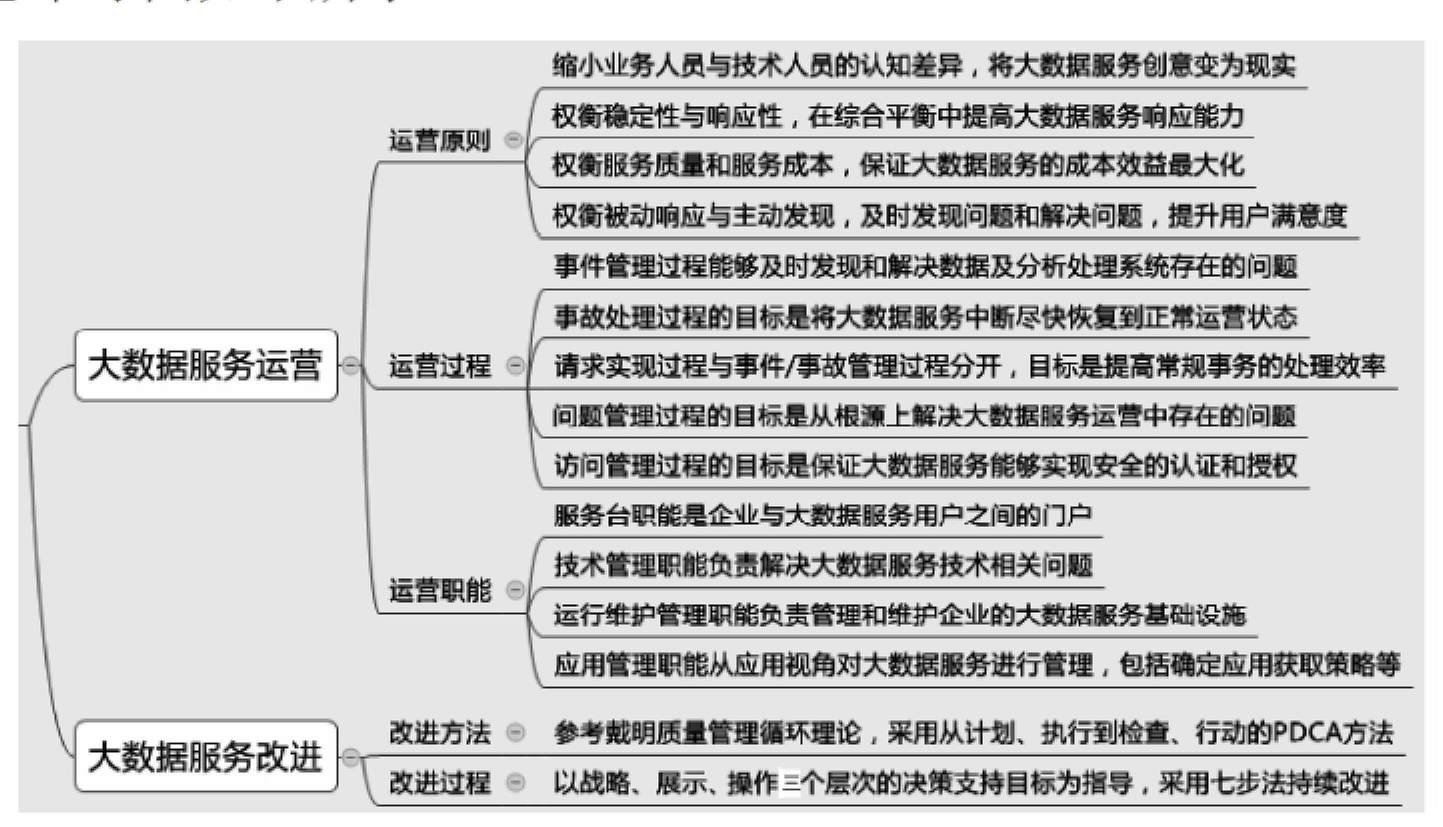
就像小宝宝在成长的过程中，有时候会生病发烧，有时候会心情不好一样，大数据服务这个刚刚出生的小宝宝，也会在运营中出现很多问题，比如突然变得不可用了，不能帮助企业实现快速的决策支持了，甚至可能会误导企业，做出错误的决策，等等。以上问题的出现都是很正常的，都需要企业及时发现问题，查找根源并找出解决问题的方法，持续地改进大数据服务。

培育大数据服务是一个持续提升的过程，包括大数据服务运营和大数据服务改进两个阶段。

在大数据服务运营阶段，需要在大数据服务运营原则的指导下，实施规范化的过程，提高发现和解决问题的效率，通过专业化职能，提高不同部门、不同岗位、不同角色的协同配合能力，为大数据服务用户提供满意的服务。

在大数据服务改进阶段，参考戴明的质量循环理论，给出了大数据服务改进的方法，然后再以决策支持在战略、战术以及操作三个层次的目标为指导，通过七步法完成大数据服务从定义到改进的过程。

本章内容思维导图如下所示：



5.1　大数据服务运营：多、快、好、省

大数据服务运营既包括事件管理、事故管理、请求实现、问题管理、访问管理等过程，又包括服务台、技术管理、应用管理等职能。

与企业面向操作的事务型应用相比，大数据服务在运营阶段会有更多问题需要解决，尤其是大数据服务属于分析型应用，许多新的需求是在运营过程中，业务人员获得新的启示后发现的。

大数据服务运营与企业运营一样，需要通过过程管理来保障。大数据服务运营过程包括事件管理、事故管理、请求实现、问题管理以及访问管理，满足特定需要的职能包括服务台、技术管理、大数据服务运维管理、大数据应用管理。

5.1.1　大数据服务运营原则

大数据服务运营原则其实就是一个综合权衡的过程。在大数据服务运营的过程中需要权衡多种因素，比如内部 IT 视角和外部业务视角、稳定性和响应性、服务质量和服务成本、被动和主动等。

1. 缩小业务人员与技术人员的认知差异

由于业务和技术人员天然上关注点不同，自然会在大数据服务运营方面存在差异，企业需要尽量消除两者之间的鸿沟。一般来讲，业务人员将 IT 作为一种满足客户和用户需求的服务、工具和手段，其更关注于大数据服务的“价值创造”，而技术人员则将 IT 看作多个不同的技术组件，其更关注的是“实现”。由于职业背景不同，理论上无法填平业务人员和技术人员之间的鸿沟，只能通过沟通和培训增强双方的理解，让双方以大数据服务价值创造为准绳来协同配合，实现企业共同的目标。

2. 权衡系统稳定性与响应性

对于大数据服务，尤其是那些与企业生产关系密切的应用，特别需要大数据服务能够

稳定运行，但是现实情况是许多因素需要大数据服务做出改变，要变就会有风险，就会影响现有应用的稳定性。当然，如果不变，又会影响对于外部市场需求的响应速度，进而影响到企业的客户的市场竞争力。因此，需要在“不变”和“变”之间取得一种平衡，在保证大数据服务稳定性的前提下取得更好的响应性。

3．权衡服务质量与服务成本

质量和成本通常是事情的一体两面，鱼和熊掌不可兼得，大数据服务的质量提升了，通常也会需要更多服务成本，企业需要在服务质量和服务成本之间进行权衡取舍。

4．权衡被动的响应与主动发现

在大数据服务运营的过程中，往往是用户发现问题后才会被动地分析和解决问题，这种被动解决问题的方式会导致用户满意度下降。如果采用实时地监控大数据服务运行的情况来主动发现其存在的潜在风险，在问题发生之前就将问题解决掉，这样就会保证大数据服务的高可用性，提升用户满意度。

5.1.2 大数据服务运营过程

大数据服务运营过程根据问题等级、问题发现方式等分为事件管理、事故管理、请求实现、问题管理、访问管理。

在大数据服务运营过程中，如果发现的问题不影响用户的正常使用，可以通过提示、告警、通知等方式处理监控事件，比如磁盘剩余空间、查询统计响应时间是否超出某个预设值等。

对于已经影响大数据服务正常使用的事件归为事故管理，比如服务器或者数据库宕机、ETL 系统故障等。可以通过事故管理过程，将待解决问题落实到专业人员，以便快速恢复大数据服务。

在大数据服务运行的过程中，也会有许多常规性的服务请求，比如新增系统账户、密码重置、数据字典修改等，这些问题纳入请求实现过程。

大数据服务应当做到按组织、用户、角色、岗位等进行授权，通过授权限定数据访问的范围和深度。可以通过事后审计，发现非正常的数据使用，尤其要注意因数据泄露引起的隐私和法律问题。

1. 大数据服务事件管理过程

事件是大数据服务在运营过程中出现并被捕获的，事件管理的目的是对大数据服务运营过程进行监控，通过监控反馈信息来发现问题，事件管理过程将这些信息进行筛选后推送到相关过程进行后续处理。

事件的来源包括大数据服务、配置项以及监控工具。大数据服务事件管理包括主动监控和被动监控两种类型。

事件管理是大数据服务运营的基础，事件管理可以发现问题后通过自动化的方式来修复问题，进而提高了大数据服务运营的效率，保证了服务运营的连续性。大数据服务运营事件处理过程如图 5-1-1 所示。

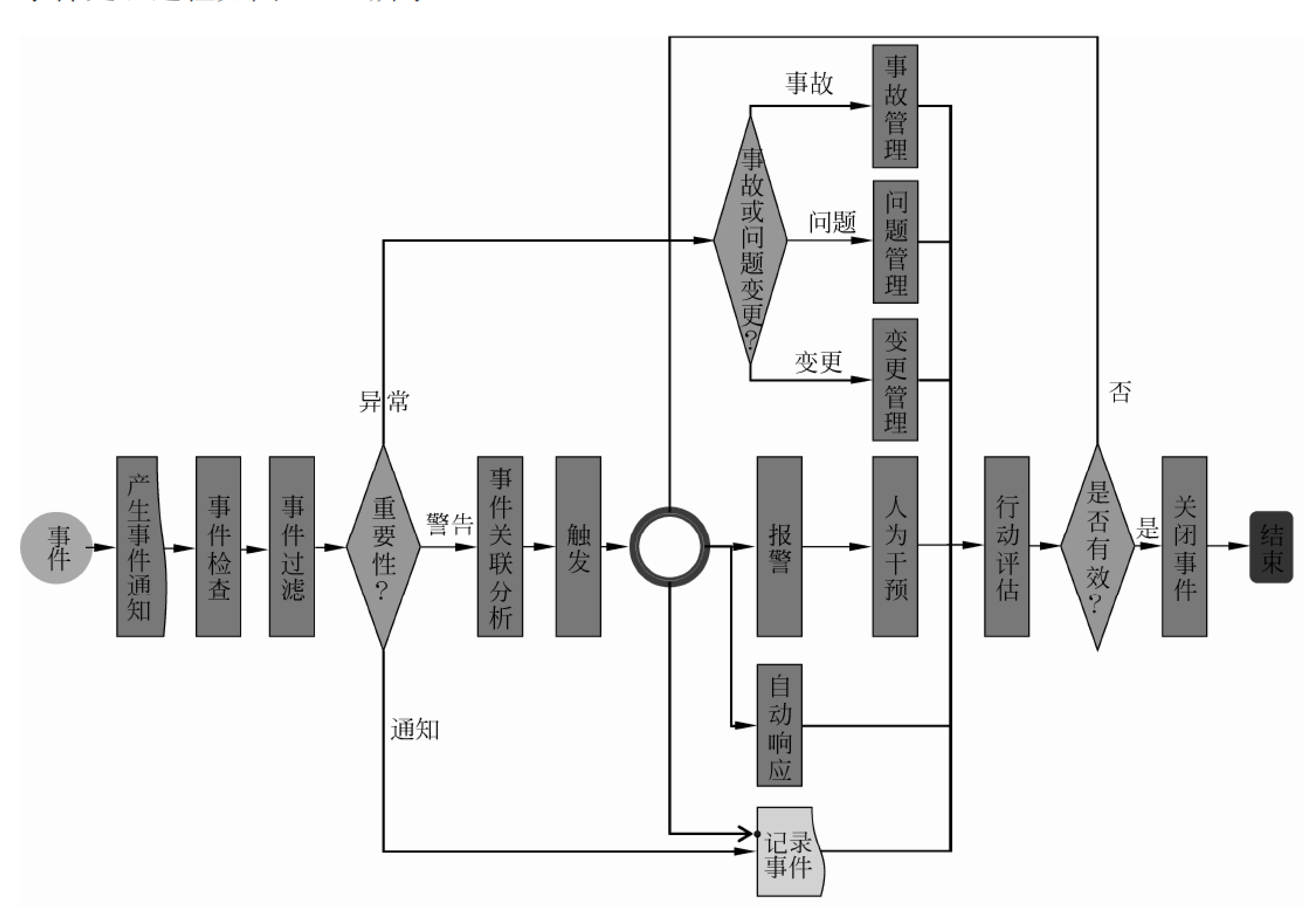


图 5-1-1　大数据服务运营事件处理过程

从图 5-1-1 可以看出，大数据服务事件处理过程为两个大的阶段。

第一个阶段为单一事件处理阶段。这个阶段主要完成大数据的分类处理，如果是通知

型事件，只需要记录事件记录即可，如果是异常事件，则需要进一步辨别后处理。

第二阶段为事件的关联分析阶段。因为一个事件的发生可能会影响其他大数据服务运营过程，如果属于比较严重的事件报警，则需要人为干预，否则需要根据事件的类型转入相应的过程来处理。事件处理完成之后，还需要进行评估，如果评估后仍然存在问题，则应当继续进行处理直至问题解决。

2. 大数据服务事故管理过程

事故不同于事件，它已经造成了大数据服务的中断，因此事故管理过程的目标是尽快恢复大数据服务到正常运营的状态。大数据服务事故管理过程如图 5-1-2 所示。

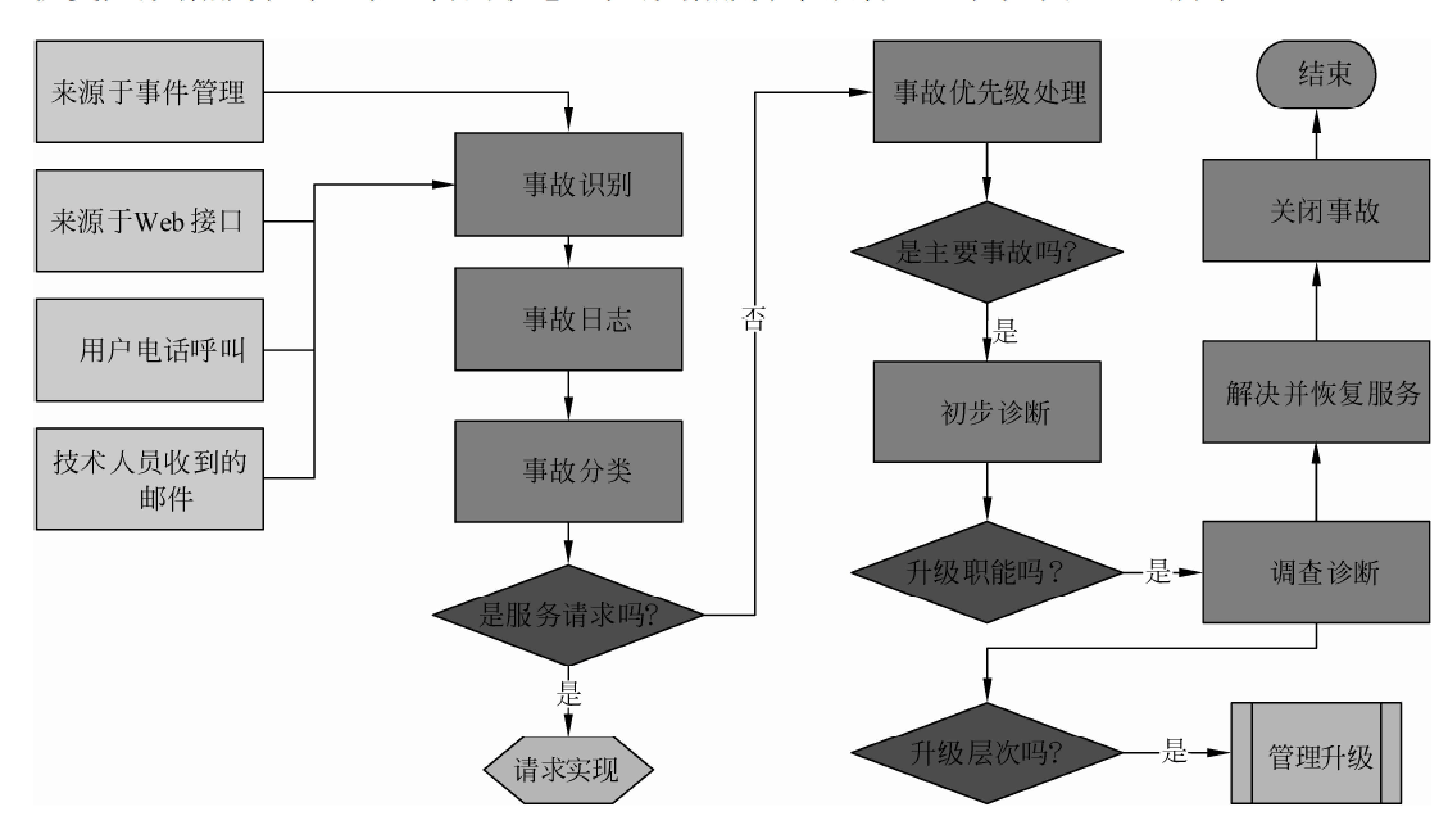


图 5-1-2 大数据服务事故管理过程

从图 5-1-2 可以看出，大数据服务事故管理过程首先接收来自不同渠道的事故源，包括电话、邮件、Web、其他事件管理过程等，然后再识别事故的级别。如果事故属于服务请求，则交给请求实现过程完成，否则需要判断事故是否为主要事故，如果是，则需要继续判断是否需要升级处理，如果需要升级处理，还需要判断是否提交到管理层处理。通过一系列的判断，最终按照时限、优先级等完成事故的处理，以保证大数据服务的快速

恢复。

3. 大数据服务请求实现过程

事件是不可以预知的，服务请求是可以计划的。可以将各种不同的服务请求进行标准化，并采用菜单形式管理服务请求，从而提高服务请求的效率。

大数据服务请求实现过程处理常规性服务，包括密码修改、联系方式修改等。

4. 大数据服务问题管理过程

大数据服务问题管理过程的目的是找出产生问题的根源，从根本上解决问题，而不是"头疼医头、脚疼医脚"。

大数据服务运营过程中产生问题的原因很多，数据质量的好坏是影响大数据服务质量最主要的因素，其次还有数据模型设计、数据展现形式、数据挖掘算法等方面。

从根本上解决大数据服务运营过程中产生的问题包括被动解决和主动解决两种方式，被动解决是在大数据服务运营阶段解决问题，而主动解决问题主要在大数据服务持续改进阶段解决。

大数据服务问题管理过程解决问题的常用方法是日志分析，大数据服务运营过程中会将用户的操作行为以及系统的运行状况记录下来，包括日志时间、操作用户、操作动作、执行结果等，操作用户可能是人也可能是系统，问题管理过程根据这些日志分析并找出出现问题的原因。

大数据服务问题管理过程需要借助配置管理过程来定位与问题有关的配置项，也可以借助知识库进行关键字查询，找到问题产生的原因以及解决的方法。此外，大数据服务问题管理过程可以作为变更管理过程的输入，通过变更管理过程来从根本上解决问题。

大数据服务问题管理过程如图 5-1-3 所示。

从图 5-1-3 可以看出，大数据服务问题管理过程就是一个主动发现问题、定位问题直至解决问题的过程。

在问题发现与判断的过程中，要对发现的问题进行分级分类，通过配置管理系统发现影响问题的环节，比如由主机引起、由中间件引起，还是由应用软件引起等，然后判断解决这一问题是否有变通方案，如果有变通方案，则执行变通方案，然后过程结束，如果没

有则将发现的问题作为错误记录到已知错误数据库，如果该问题再次发生，则可以作为经验快速找到解决问题的方法。然后判断是否需要变更，如果不需要则直接解决问题，比如只需重新启动问题相关的几个中间件服务，如果需要变更，则需要进入大数据服务问题变更过程。

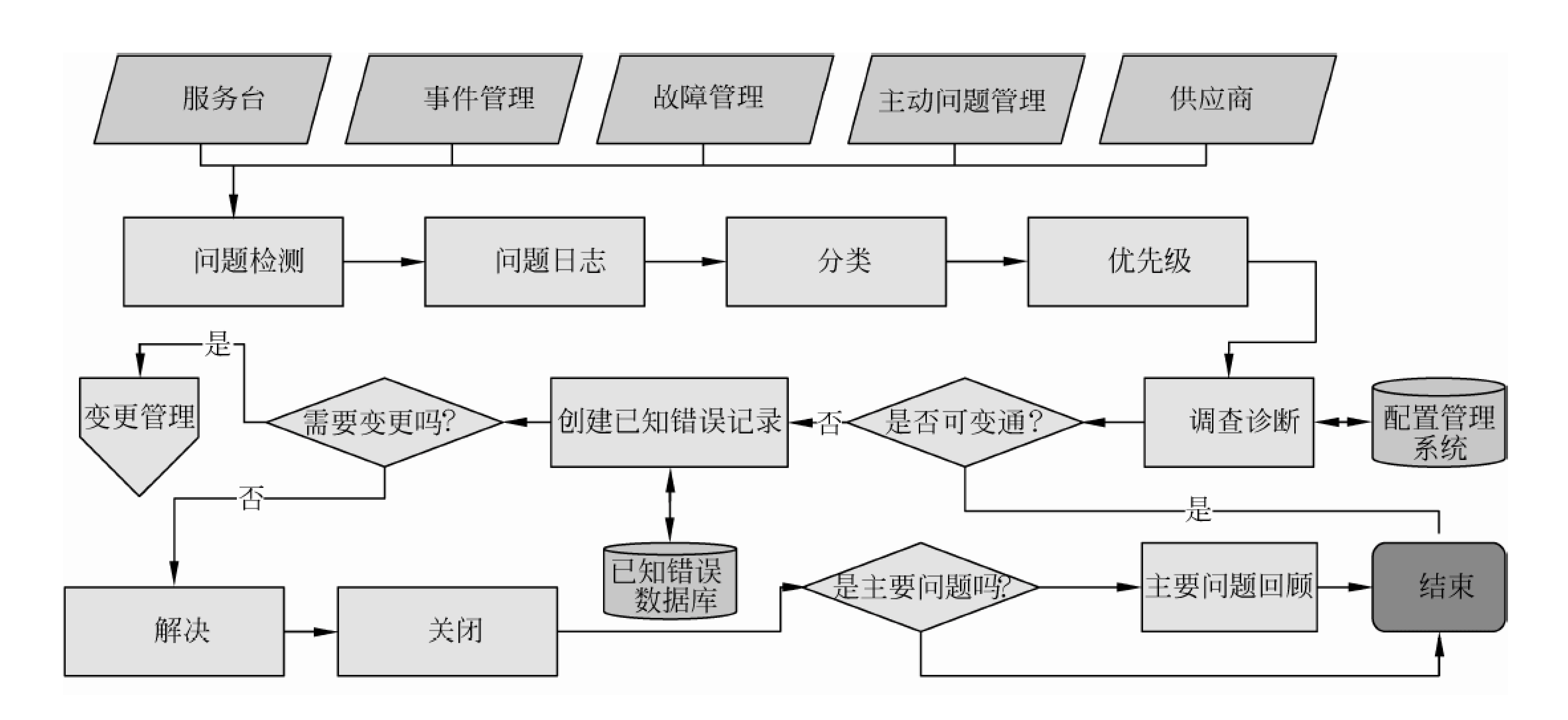


图 5-1-3　大数据服务问题管理过程

5. 大数据服务访问管理过程

大数据服务访问管理过程的目标是使得合适的人能够获得授权的大数据服务，因此需要账户管理、认证管理以及授权管理。此外，为了发现大数据服务运营过程中存在的潜在风险或者找出出现问题的责任方，需要对大数据服务使用日志进行审计。

5.1.3　大数据服务运营职能

大数据服务运营组织负责日常事务的处理，保证大数据服务的正常运营。大数据服务运营的职能包括服务台、技术管理、运行维护管理以及应用管理。

大数据有了，但是它需要有效地运营才能够发挥应有的作用，包括建立什么样的组织、需要什么样角色的人员参与，这些人员或角色之间如何协调配合结合大数据使用中存在的

问题等，这些都属于大数据运维的范围。

大数据运维的数据形式包括结构化和非结构化两种类型，数据来源于组织内部和组织外部。大数据运营架构包括数据的采集与整合、元数据管理、数据集市、数据挖掘，强调对数据的管理与利用。大数据服务实现的形式包括数据服务、统计报表、趋势分析。

1. 大数据服务服务台职能

服务台是大数据服务的用户与企业交流的门户，用户可以借助服务台反馈大数据服务使用过程中存在的问题，或者提出服务请求等，服务台也应当及时地将发现的事件、事故、问题等通知给用户，最终构建一个用户与大数据服务平台之间沟通的桥梁。

2. 大数据服务技术管理职能

大数据服务技术管理职能解决与技术相关的问题。

大数据服务的运营不但需要电力、空调、照明等机房基础设施，而且需要主机、网络、存储、中间件、数据库等系统硬件和系统软件的支持，如果大数据服务出现问题，则需要不同方向的技术专家参与，技术管理是大数据服务的重要保障。

大数据服务技术管理职能分为知识经验提供角色和资源供给角色。

3. 大数据服务运行维护管理职能

与大数据服务技术管理职能相比，大数据服务运行维护管理职能负责管理和维护企业的大数据服务基础设施，以保证能够交付业务所需的大数据服务。

大数据服务运行维护职能具体包括运营控制、设施管理。运营控制包括控制台管理、工作计划、备份与恢复、打印和输出。实施管理包括电力、空调、空间等管理。

大数据服务运行维护管理职能的角色包括运营控制员和基础设施管理员。

4. 大数据服务应用管理职能

大数据服务应用管理职能是从应用视角进行大数据服务管理的。应用管理职能需要确定应用获取的策略，比如，购买应用还是自主研发应用？大数据服务应用管理职能需要对大数据服务全生命周期进行管理，包括需求、设计、构建、部署、上线运营、优化。

大数据服务应用管理职能的角色包括应用负责人、应用分析师、应用架构师。

5.2 大数据服务改进：自强不息止于至善

大数据服务不是一蹴而就的，是需要一个不断改进完善的过程，发现问题和差距并持续改进是提升企业决策能力的唯一途径。

大数据服务为了满足市场需求，需要不断地优化完善才行。可以通过对大数据服务进行评估和对标，发现大数据服务与决策支持目标之间存在的差距，然后再进行改进完善。

5.2.1 大数据服务改进方法

按照戴明的质量管理循环理论——计划-执行-检查-行动（Plan-Do-Check-Action，PDCA）大数据服务的改进也同样采用计划-执行-检查-行动的方法。大数据服务改进模型如图 5-2-1 所示。

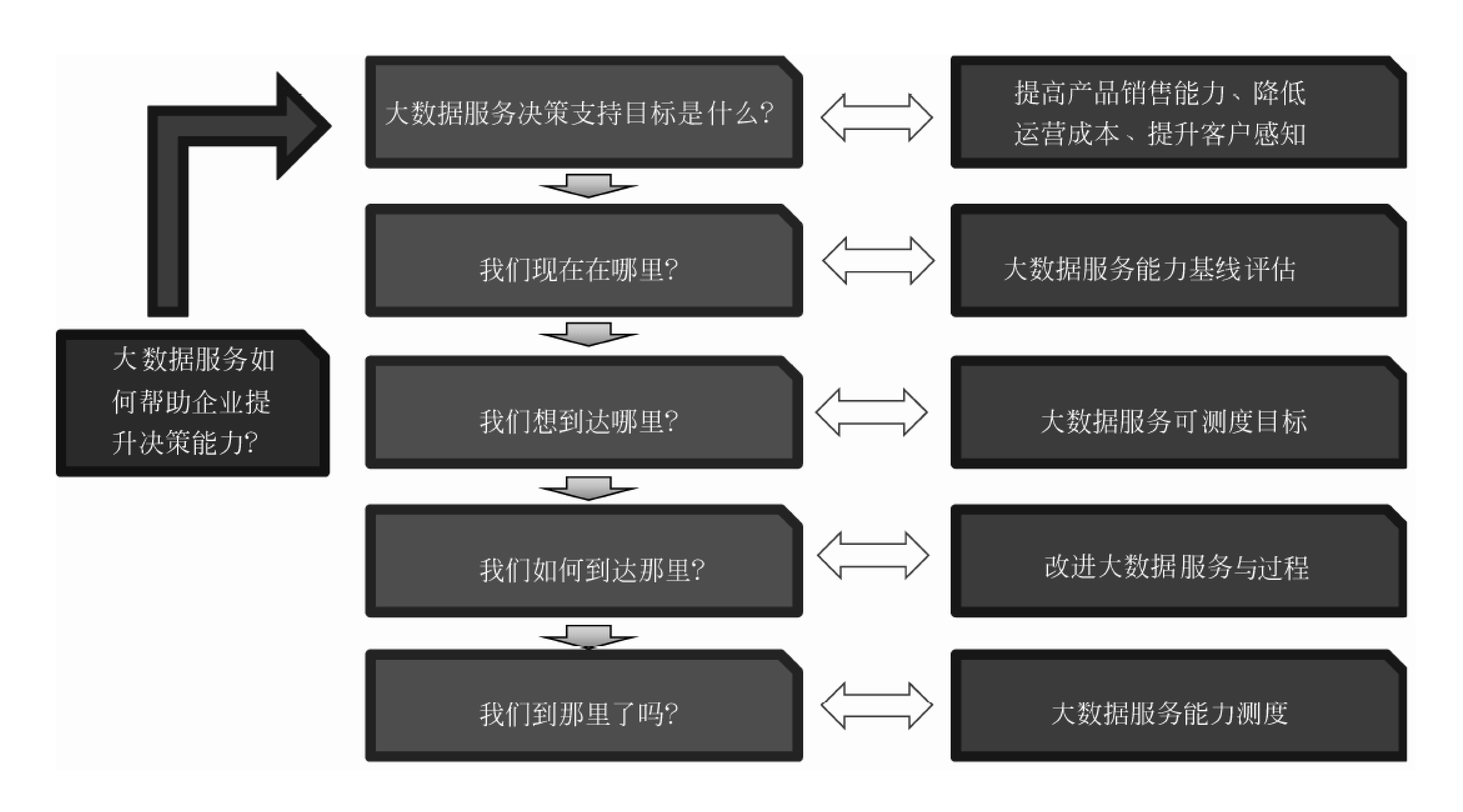


图 5-2-1 大数据服务改进模型

从图 5-2-1 可以看出，大数据服务改进模型首先是确定决策支持的目标，包括产品销售能力提升目标、运营成本降低目标或者客户感知提升目标。然后，企业需要对大数据服务能力进行基线评估，确定大数据服务的能力现状。接着，基于大数据服务在产品销售能力提升、运营成本降低或者客户感知提升方面的总体目标，确定可测度的 SMART（具体、可测度、可实现、相关性、截止时间）目标。再然后，结合大数据服务能力现状和提升目标，确定大数据服务的能力差距并实施改进。最后，对改进成果进行评估，确认大数据服务能力提升的程度。

5.2.2　大数据服务改进过程

大数据服务改进过程通常可以采用七步法来完成，如图 5-2-2 所示。

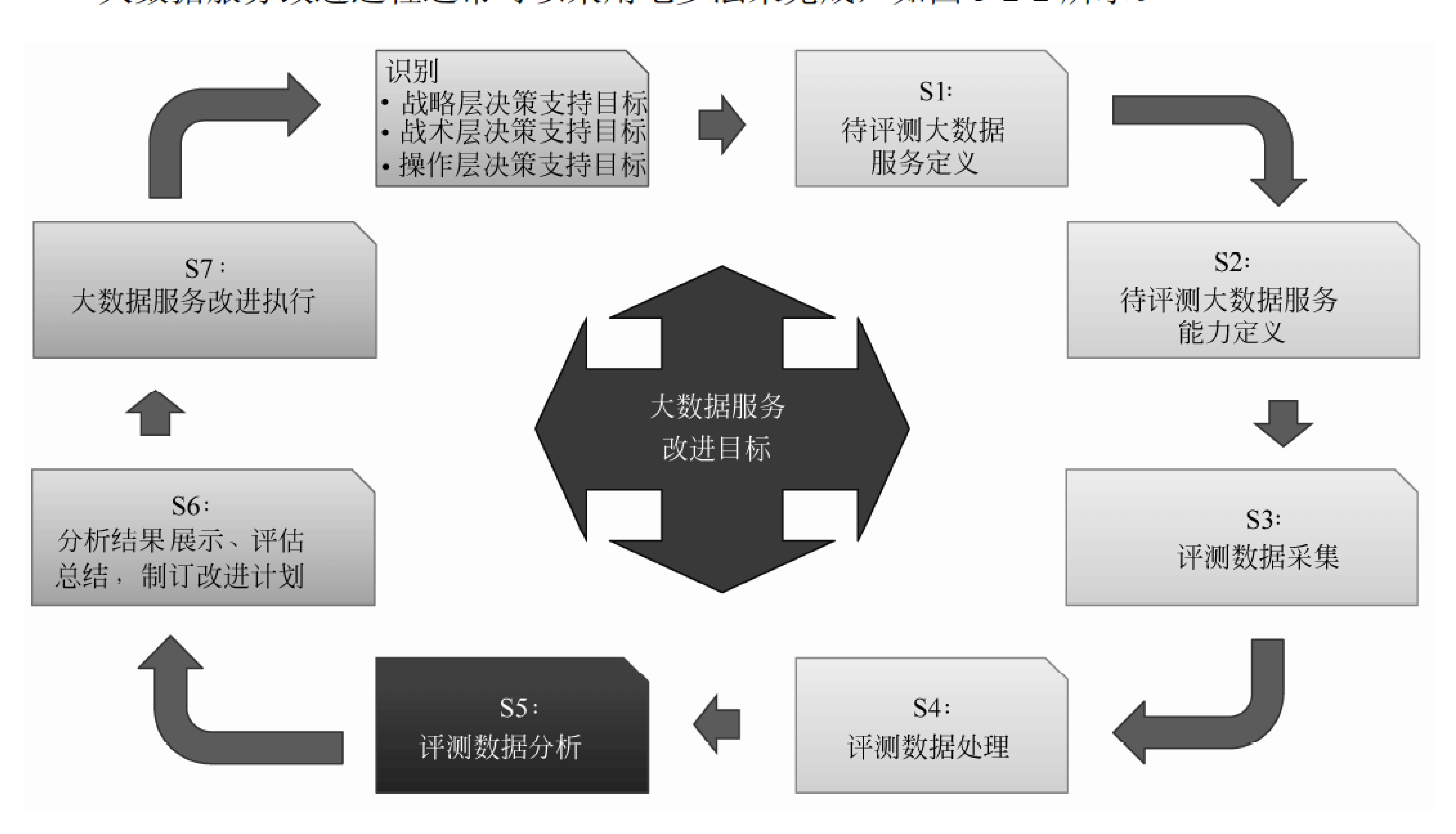


图 5-2-2　大数据服务改进过程（七步法）

从图 5-2-2 可以看出，大数据服务改进过程以大数据服务改进目标为指导，基于企业战略、战术以及操作三个层次的决策支持目标，需要完成 7 个步骤，形成一个从待评测大数据服务定义到大数据服务改进执行的闭环过程。

5.3 主要内容回顾

进入运营阶段的大数据服务，需要经过不断地调整、优化，才能体现其在企业生产经营决策中的重要作用。

大数据服务运营一方面要满足业务人员对于大数据服务的正常使用需要，另一方面还要对其进行优化完成，提升其对企业生产经营决策的业务价值。

为了使得大数据服务满足企业正常的使用需要，需要构建大数据服务管理体系，通过设置多个相互配合的运营过程和职能，保障企业能够快速、高效地发现和解决大数据服务运营过程中出现的问题。

为了提升大数据服务在企业生产经营过程中的决策价值，需要从业务角度出发，设定战略、战术以及操作层次的决策目标，然后检查大数据服务与满足这些决策目标的差距，找出缩小这一差距的方法，通过变更或者新增大数据服务，提升大数据服务的决策支持能力。

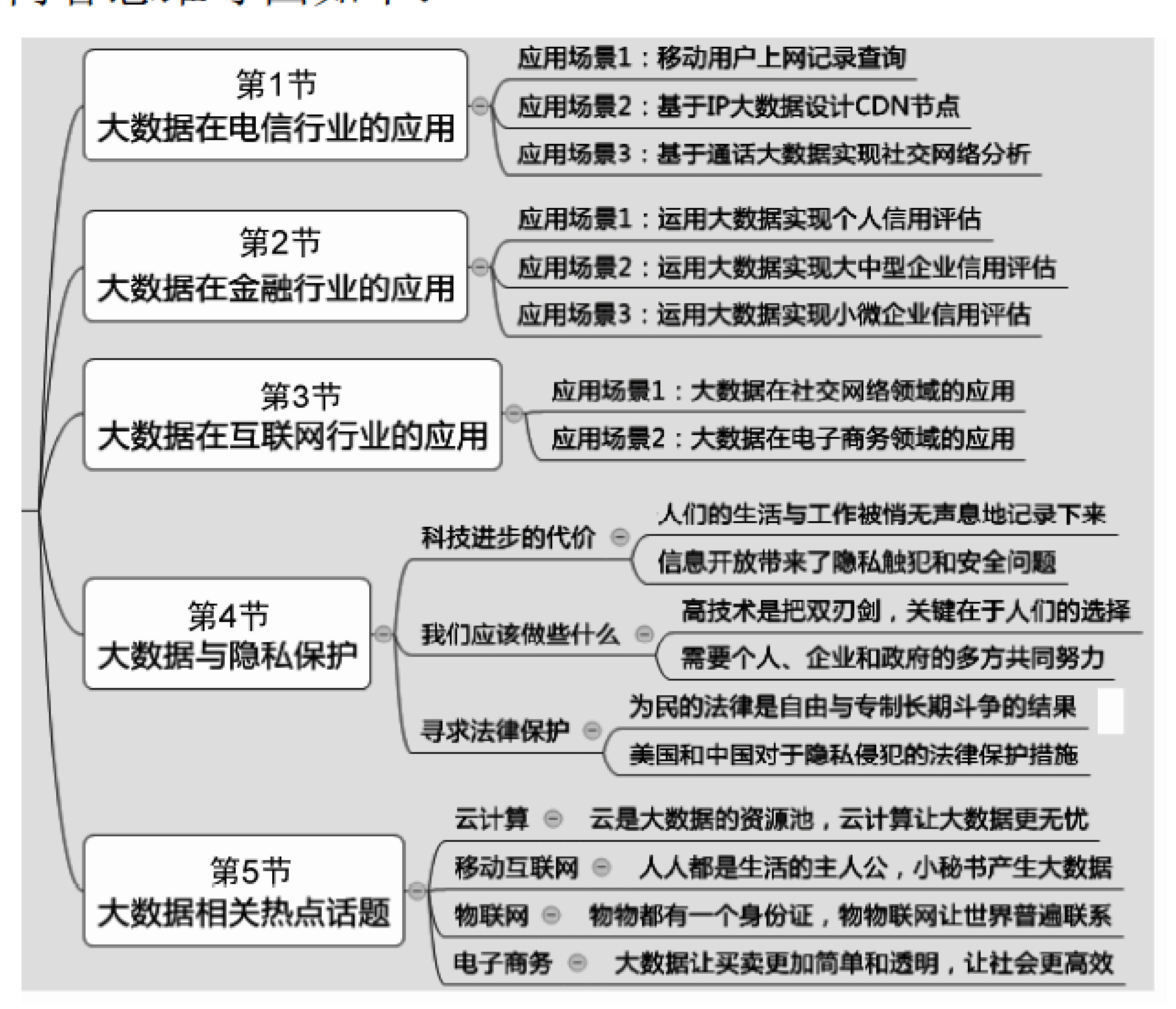
第 6 章 腾飞：在实践中检验真理

“不养儿不知父母恩”，家庭对孩子的“培育”是非常辛苦的，何况父母对于孩子总是有更高的期望。“一分耕耘一分收获，十分耕耘十分收获”，父母对孩子的精心培育，一定会有好的回报的。企业对待大数据服务，就像父母对待“孩子”一样，通过不断地优化和完善，大数据服务这个“孩子”也一定会不负众望，助力企业大展宏图，像巨龙一样“腾飞”起来。

实践是检验真理的唯一标准。大数据服务能否帮助企业取得成功，还需要在行业应用实践中得到答案。下面以电信、金融和互联网三个行业为例，说明大数据服务如何在企业生产经营实践中应用。

大数据服务要助力企业实现“腾飞”，必然要开放数据，而开放数据就可能会引起个人隐私触犯，商业秘密泄露，个人、企业或者社会安全受到威胁，以及触犯国家法律法规等一系列问题，这些都是企业在应用大数据服务的过程中特别需要注意和防范的。

此外，与大数据相关的热点话题很多，比如云计算、移动互联网、物联网、电子商务等，如果企业能够掌握这些社会热点与大数据的密切联系，将能够让大数据服务“飞”得更高、更远。本章内容思维导图如下：



6.1 大数据在电信行业的应用

通信大数据既包含真实可靠的用户属性信息，又包括通话、上网等用户实时行为信息，可以反映个体与群体的社交关系、需求偏好、行为特征等。

如今的电信运营商，在腾讯、阿里、360 等厂商在电信网络之上提供服务（On The Top，OTT）和京东、蜗牛移动、国美等虚拟运营商的多面夹击下，处于日益激烈的市场竞争之中，逐渐失去以往那种依靠经营垄断生存的好日子，在新的市场环境下，必须重新思考，寻找新的业务创新点。

随着移动通信网络和移动智能终端的飞速发展，预示着移动互联网和大数据时代的到来，在新的形势下，电信运营商需要充分发挥以通信网络为核心的大数据优势，提升战略管理水平与运营能力。

对于电信运营商来说，其优势主要体现在三个方面。

第一，电信运营商具有庞大的通信网络资源和海量的信息通信记录，可以实时掌握用户的通信行为和地理位置，比如通话行为、上网行为、移动轨迹等；

第二，电信运营商具有亿级的庞大用户群，并且用户信息大多采用实名认证方式，基本真实可靠；

第三，电信运营商具有完善的渠道体系和庞大的营销与服务渠道资源，拥有线上和线下资源协同优势。

电信运营商的三大优势为大数据运营提供了很好的数据基础，在电信产品同质化的今天，电信运营商应当发挥在大数据方面的差异化优势，充分挖掘和释放信息通信大数据的潜力，用于企业的发展战略、建设、运营等各个环节之中，同时，应当积极与其他行业合作，推动满足全社会需要的大数据应用创新。

下面就以电信运营商已经实现的或者可以实现的两个大数据应用场景为例，分析电信大数据实现应用创新的方法和思路。

6.1.1 应用场景 1：移动用户上网记录查询

1. 问题的产生

第一代移动通信主要解决了人与人之间的电话通信问题，第二代移动通信则使得人们可以用手机上网了。但由于第二代移动通信的上网速率很低，人们只能通过 WAP 方式变相满足手机上网需求，由于网速限制，WAP 手机上网方式仅能满足简单的网页浏览等功能。

为了提升移动用户上网速度， 3GPP（the 3rd Generation Partnerships Project）提出了发展第三代移动通信（简称 3G）的计划。借助 3G 移动通信网络，上网速度可以达到几十兆，大大提升了移动用户访问互联网的速度和质量。2008 年 12 月，工信部为我国三大运营商发放了 3G 牌照，标志着我国开始全面迈入 3G 时代。

3G 网络虽然上网速度快，但受限于无线频谱资源，需要采用基于流量收费的方式，在 3G 网络应用的早期，应用和用户都很少，还没有出现数据存储和流量查询问题。后来，随着移动互联网应用的飞速发展，移动用户形成的上网记录快速增加，每天就有 PB 级别的数据量，在这种情况下，由于传统的关系型数据库无法实现 IT 基础设施资源的横向扩展，因此无法满足用户对上网记录的快速查询需求。

因此，当移动用户对上网费用提出质疑并且要求核实移动上网数据时，由于电信运营商无法提供及时、准确的上网记录清单，许多情况下不得不采用向用户退费的简单处理方式，这对电信运营商开展移动上网业务造成了许多负面影响，并且导致了企业大量的收入流失。

为了解决这一问题，迫切需要电信运营商采用先进的、满足用户对移动上网记录快速查询需求的解决方案。

2. 问题分析与方案设计

通过观察发现，移动用户上网记录的主要特征是数据量太大并且数据产生的速度快，随着新的上网记录不断产生，传统的关系型数据库存取性能急剧下滑，甚至经常出现查询后“死机”的现象，根本无法满足查询需求。

究其根源，传统的关系型数据库产生于事务型应用盛行的时代，通常采用“主机+磁

盘阵列”的集群架构，主机集群可以满足多用户高并发的性能需求，磁盘阵列则保证数据存放的安全性。当面向 GB 或者 TB 级别的数据库查询时，系统的响应时间通常可以提高到几秒之内，还是可以满足需求的。如果查询性能下降，一般可以通过横向扩展主机的方式，或者通过纵向提升主机配置的方式来解决。

随着移动数据业务的飞速发展，移动用户上网记录数达到每天 PB 级别的数量级，那么采用传统关系型数据库的架构方式就难以满足要求了。

由开源组织阿帕奇发布的 HBase 是一款分布式列式数据库，对于主机的要求不高，采用普通 PC 服务器即可。由于 HBase 数据库具有良好的横向扩展能力，并且系统整体性能随着服务器资源的增加可以实现线性提升。如果把电信运营商的移动用户上网记录存储到 HBase 数据库集群中，既能使用开源数据库节约成本，又能够满足移动用户上网记录大数据的查询性能要求。

为了不影响现有生产系统（采集系统、计费系统）的正常运行，电信运营商采用在 GGSN 和 SGSN 两个网关之间部署探针（分光器）的方式，实现对移动用户上网记录的采集，然后再通过各个探针采集的数据逐级汇聚到 HBase 集群中。基于 Hadoop/HBase 实现对移动用户上网记录采集的解决方案如图 6-1-1 所示。

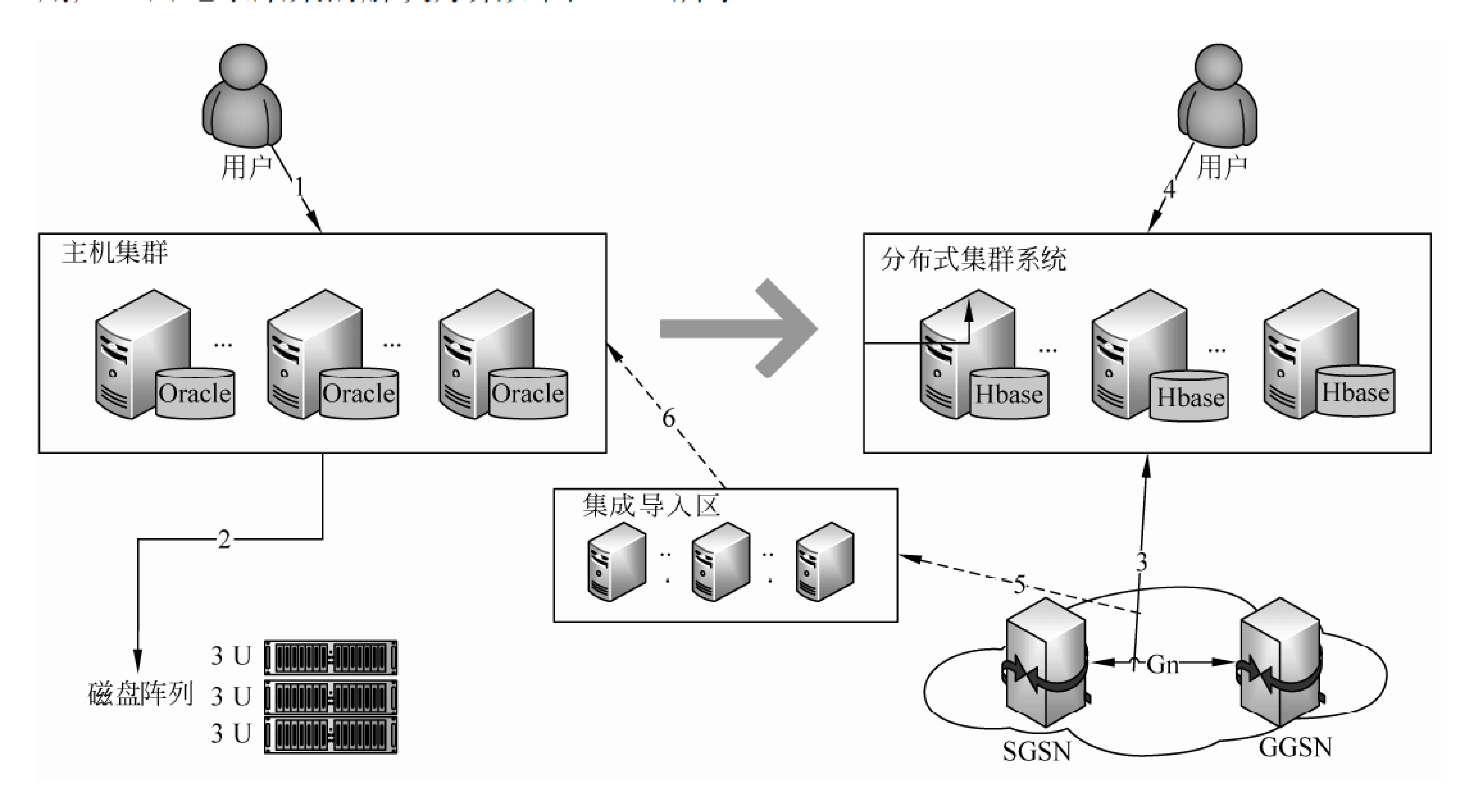


图 6-1-1　关系型数据库集群架构到分布式集群架构的转变

HBase 数据库虽然能够满足用户按照时间段和手机号码查询上网记录明细的需求，但是在多维度的统计分析方面能力非常有限。为了实现按时间段、按地域、按网络类型等维度的统计，需要分布式数据库 HBase 与传统关系型数据库相互配合，HBase 负责提供少量关键字的移动上网记录明细查询功能，比如按照用户手机号码、目标 IP、网址 URL 等条件查询，而原始的移动用户上网记录通过汇总形成大颗粒度的统计数据后进入关系型数据库。关系型数据库可以通过分区、索引、中间表等方式提高统计分析的效率。

通过采用开源分布式数据库架构方式，解决了移动用户上网记录大数据存取效率的问题，为电信运营商的业务部门和通信用户提供了关于移动上网流量的真实凭证，节约了电信运营商的 IT 投资，提升了用户感知，最终提升了电信运营商的对外形象和整体竞争能力。

6.1.2　应用场景 2：基于 IP 大数据设置内容交付网络节点

1. 问题的产生

内容分发网络（Content Delivery Network，CDN），是建立于现有互联网基础之上的一层智能虚拟网络，其通过将用户的请求重新导向离用户最近的 CDN 节点，使用户可就近取得所需内容，解决 Internet 网络拥挤的状况，提高用户访问网站的响应速度。

随着移动互联网时代的到来，“平台+应用”的商业模式加快了应用创新的步伐，应用的数量不断增多。由于同一应用不一定在所有区域部署，因此用户可能需要跨越不同省份、不同电信运营商网络才能到达应用。在移动用户到应用之间的网络路径中，跨越电信运营商的网关带宽有限性成为跨网访问的瓶颈，降低了移动用户的上网速度和客户感知水平。

因此，在移动互联网中搭建 CDN 是应用发展的需要，也是提高用户应用访问速度的有效手段。2009 年 10 月，CDN 服务提供商网宿科技在深交所上市，2010 年 10 月，CDN 服务提供商蓝汛（ChinaCache）在美国纳斯达克证券交易所上市，标志着 CDN 巨大的发展潜力。

CDN 服务提供商 CDN 节点部署的原则是实现移动用户对应用的就近访问。但是现有的 CDN 服务提供商无法获知访问应用的用户归属地，进而无法准确地在用户请求多的用

户归属地为用户访问的应用设置 CDN 节点。

随着 CDN 服务提供商与电信运营商合作力度加大，未来电信运营商和 CDN 服务商合作共同在移动互联网中搭建 CDN 网络成为趋势，如果能合理设置 CDN 节点，就能够帮助 CDN 解决方案提供商为用户提供更加快捷的网络访问体验，也可提高运营商的应用价值。合理设置 CDN 节点可以实现包括移动用户、电信运营商、CDN 服务提供商在内的各个参与方的共赢，如图 6-1-2 所示。

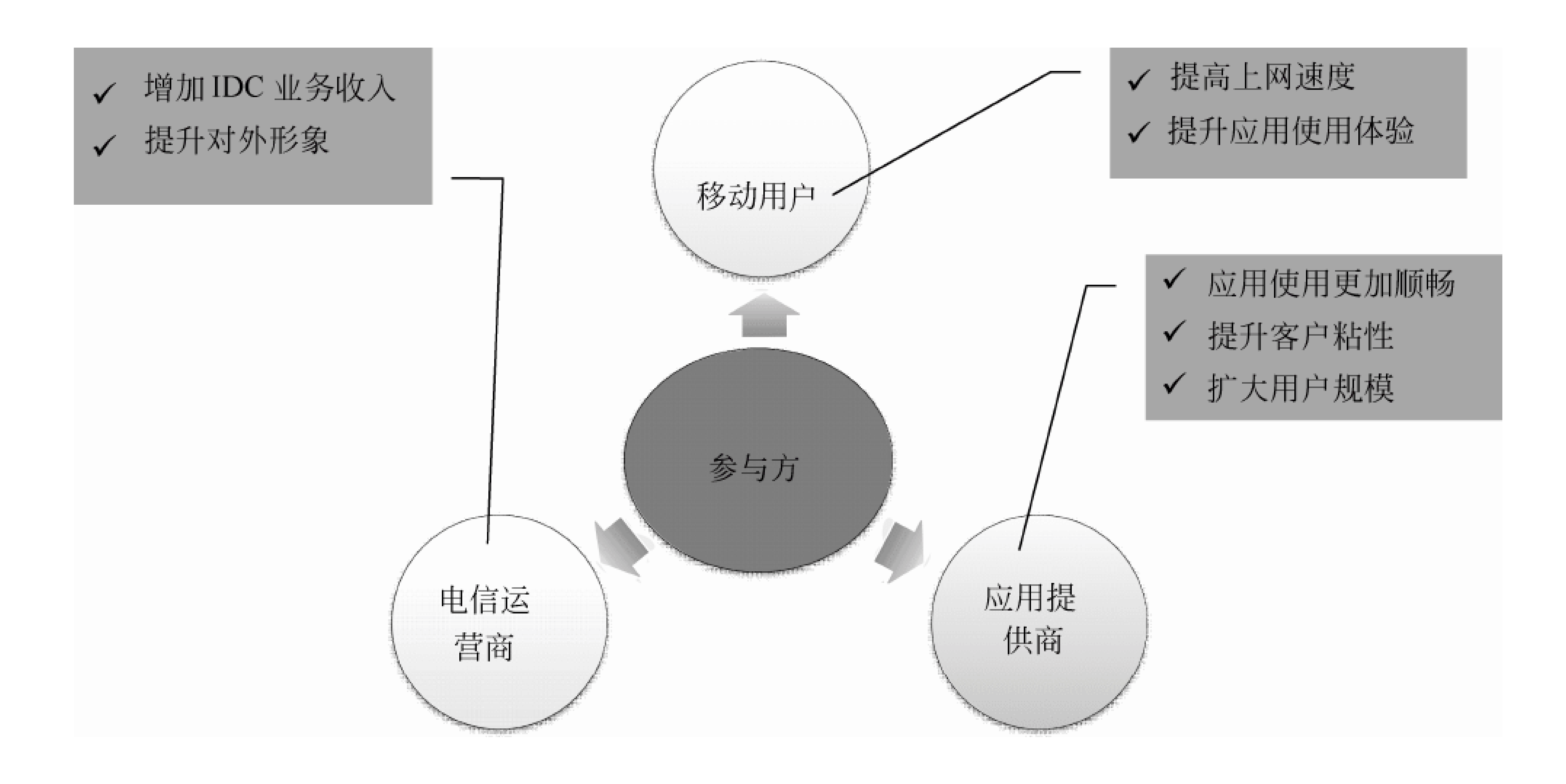


图 6-1-2　合理设置 CDN 节点对各个参与方的价值

从图 6-1-2 可以看出，合理设置 CDN 节点，首当其冲的受益者为移动用户，同时可以增加电信运营商的收入，增强应用提供商的客户黏性，扩大应用的使用者规模。因此，在满足各方需要的情况下，如何部署 CDN 节点成为一个亟待解决的问题。

2．问题分析与方案设计

内容交付网络（CDN）通过在网络边缘节点部署应用，实现用户对应用的就近访问，从而提高了移动用户应用访问速度。移动用户到应用之间的网络路径越长，比如跨省或者跨越多个运营商的通信网络，那么网络延时也就越长，如果网络路径需要穿越电信运营商之间的互通网关，那么互通网关就会成为影响上网速度的最大瓶颈，如图 6-1-3 所示。

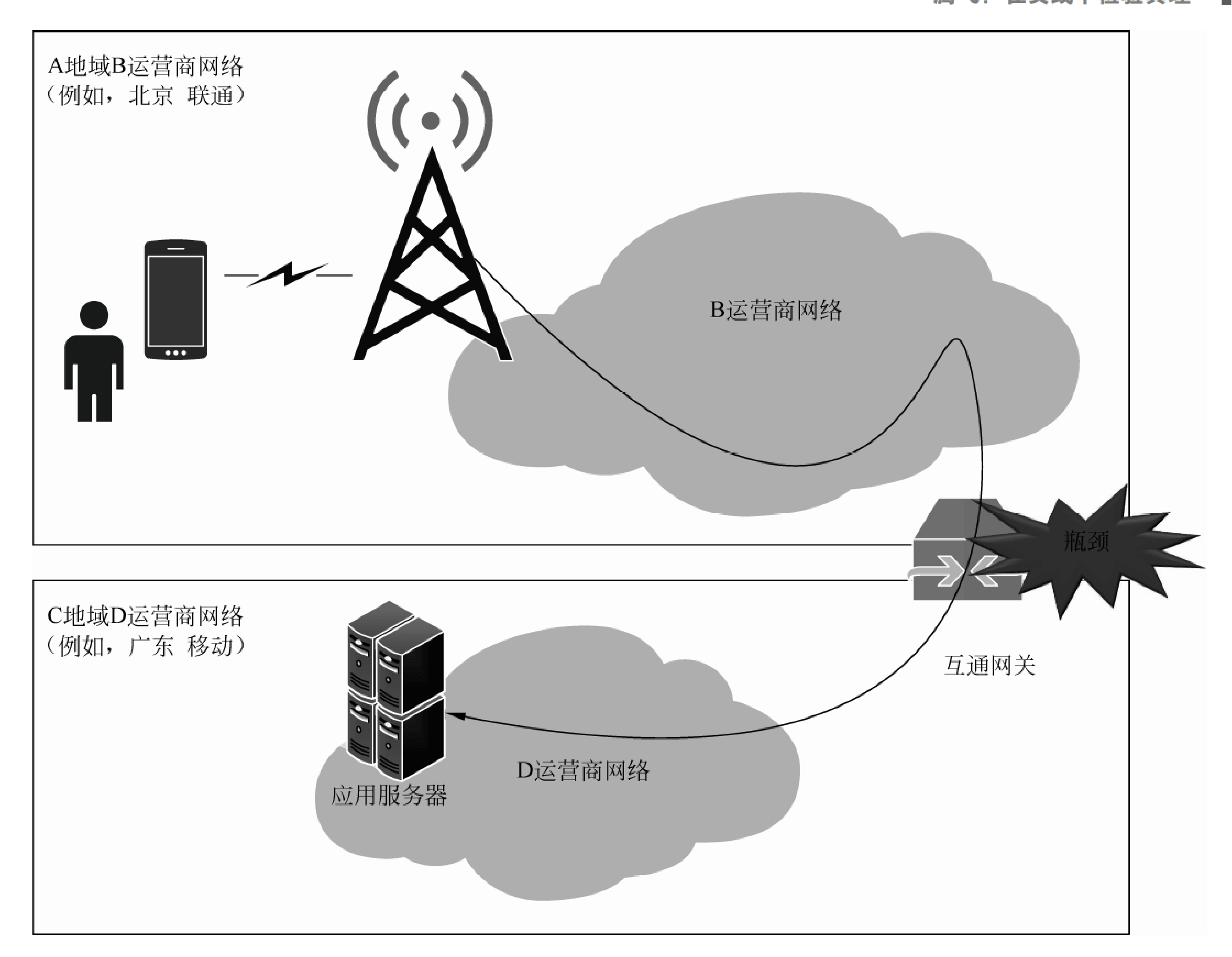


图 6-1-3 互通网关成为提升移动用户上网速度的瓶颈

因为由于竞争关系，电信运营商之间的互通网关的数量和带宽是非常有限的，因此提升移动用户应用访问速度的方法就是发现移动用户是否跨地域、跨互通网关访问应用，如果是，则建议应用提供商在移动用户归属地的电信运营商网络内增加 CDN 节点。当然，还需要对应用价值和移动用户价值高低进行评估，如果是整体价值高，则建议增加 CDN 节点。通过合理设置 CDN 节点解决问题的总体思路如图 6-1-4 所示。

移动用户上网记录是移动用户访问应用时产生的，移动用户访问一个应用或者网页，都会记录多条上网记录，累计下来每天都会产生 PB 级别、数百亿条的上网记录数据。

上网记录内容包括移动终端、网络、应用三个方面的信息。在移动终端侧，包括 IMEI（移动终端标识）、IMSI（移动用户标识）、MSISDN（移动用户号码）、移动终端 IP 地址等信息；在通信网络侧，包括上网时间、上行数据流量、下行数据流量、上网时长、网络类型、位置区域代码等信息；在应用侧，包括应用部署 IP 地址、URL（网址）等。

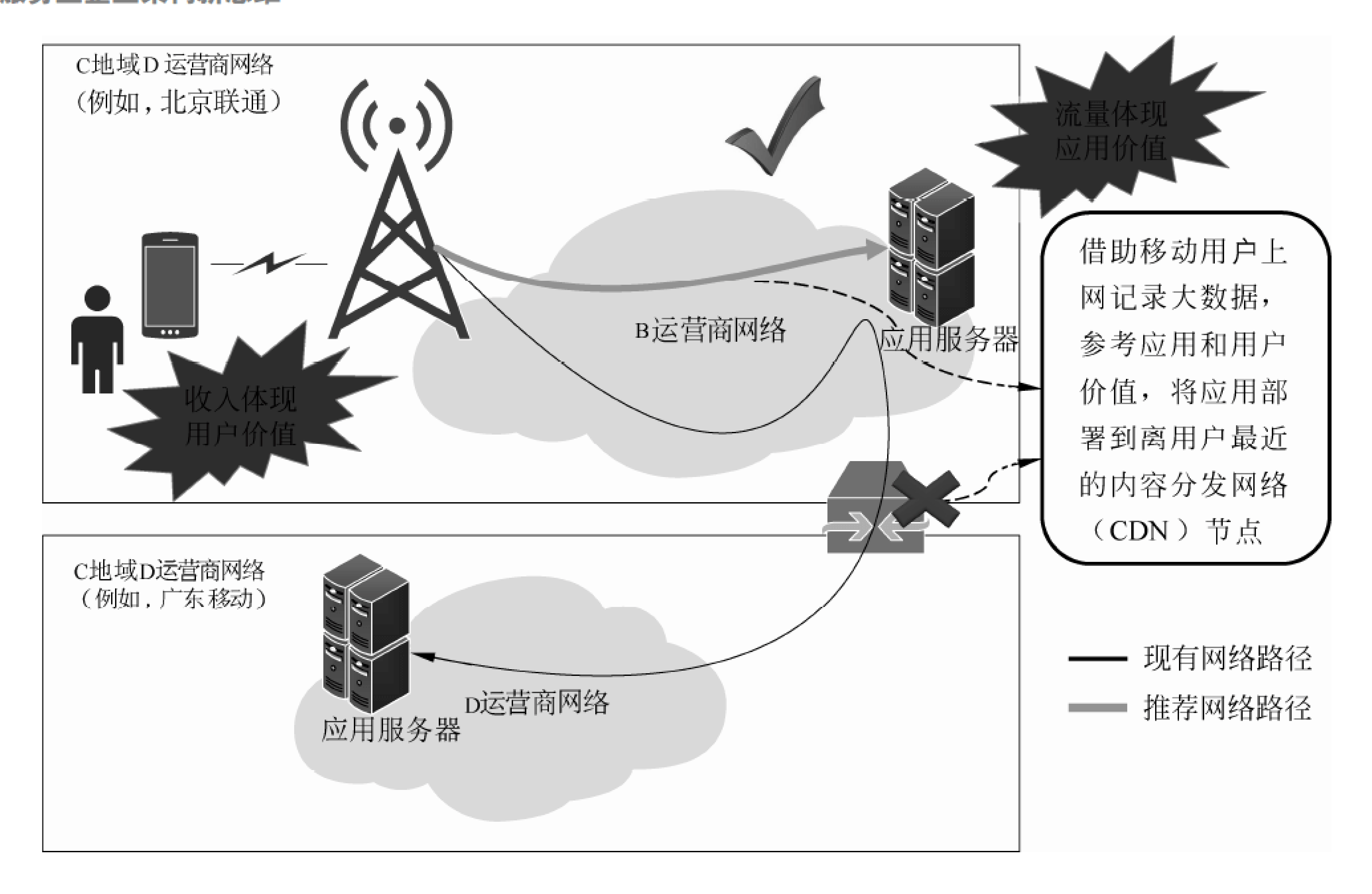
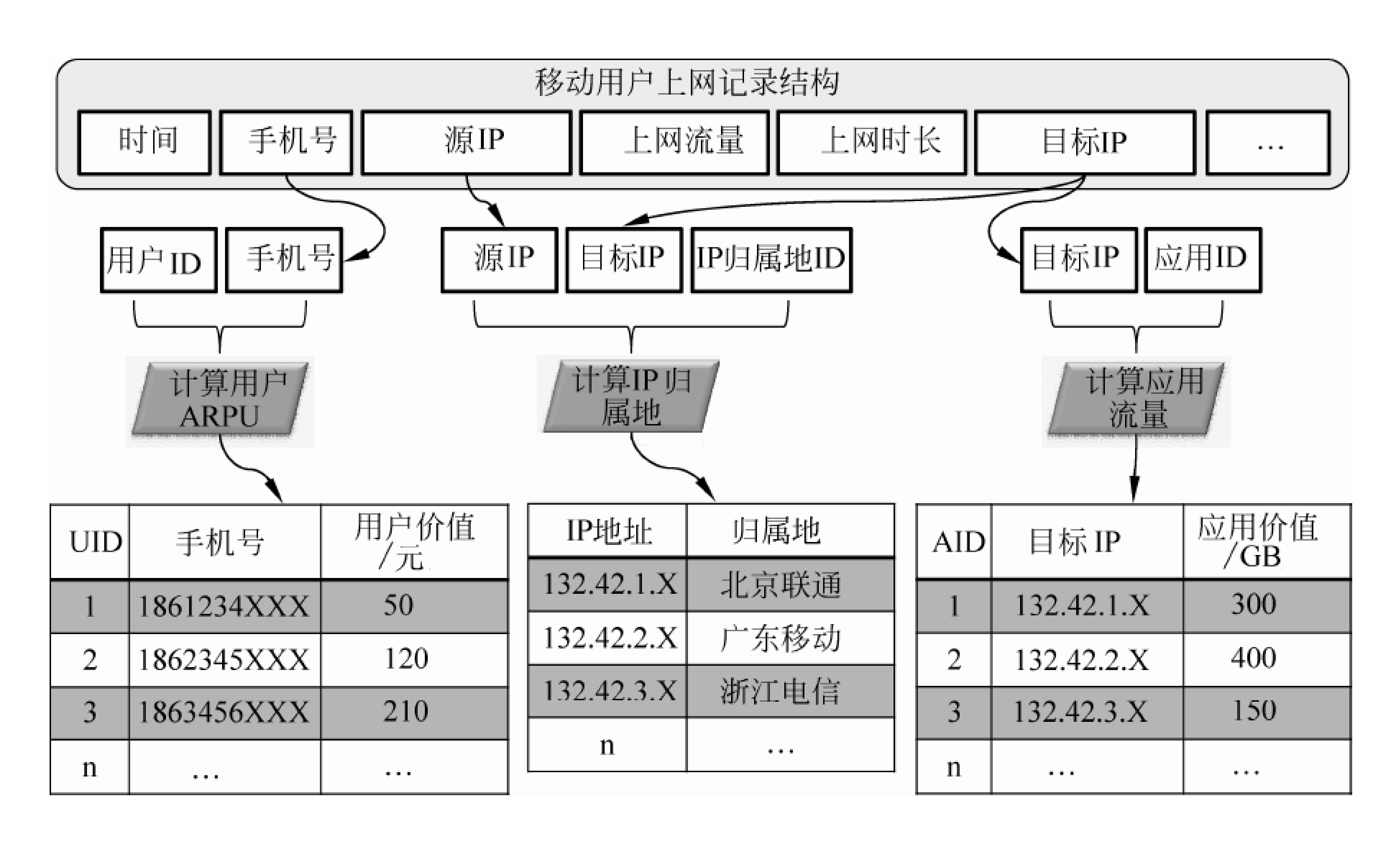


图 6-1-4　通过设置 CDN 节点提升移动用户上网速度的方法

如果以用户上网记录大数据为基础，辅以关联数据，就会计算出用户价值、应用价值以及网络访问路径，计算方法如图 6-1-5 所示。



UID	手机号	用户价值/元
1	1861234XXX	50
2	1862345XXX	120
3	1863456XXX	210
n	…	…

IP地址	归属地
132.42.1.X	北京联通
132.42.2.X	广东移动
132.42.3.X	浙江电信
n	…

AID	目标 IP	应用价值/GB
1	132.42.1.X	300
2	132.42.2.X	400
3	132.42.3.X	150
n	…	…

图 6-1-5　用户和应用的价值/归属地的计算方法

其中，用户 ARPU=流量×时长×资费，应用流量=Σ目标 IP 对应流量。

完成用户和应用的价值/归属地的计算后，还需要从价值和归属地维度进行排名。用户归属地与应用归属地不一致，才会形成设置新的 CDN 节点的需求。用户和应用价值越高，设置新的 CDN 节点才更有意义。

不同角度的价值排名如图 6-1-6 所示。

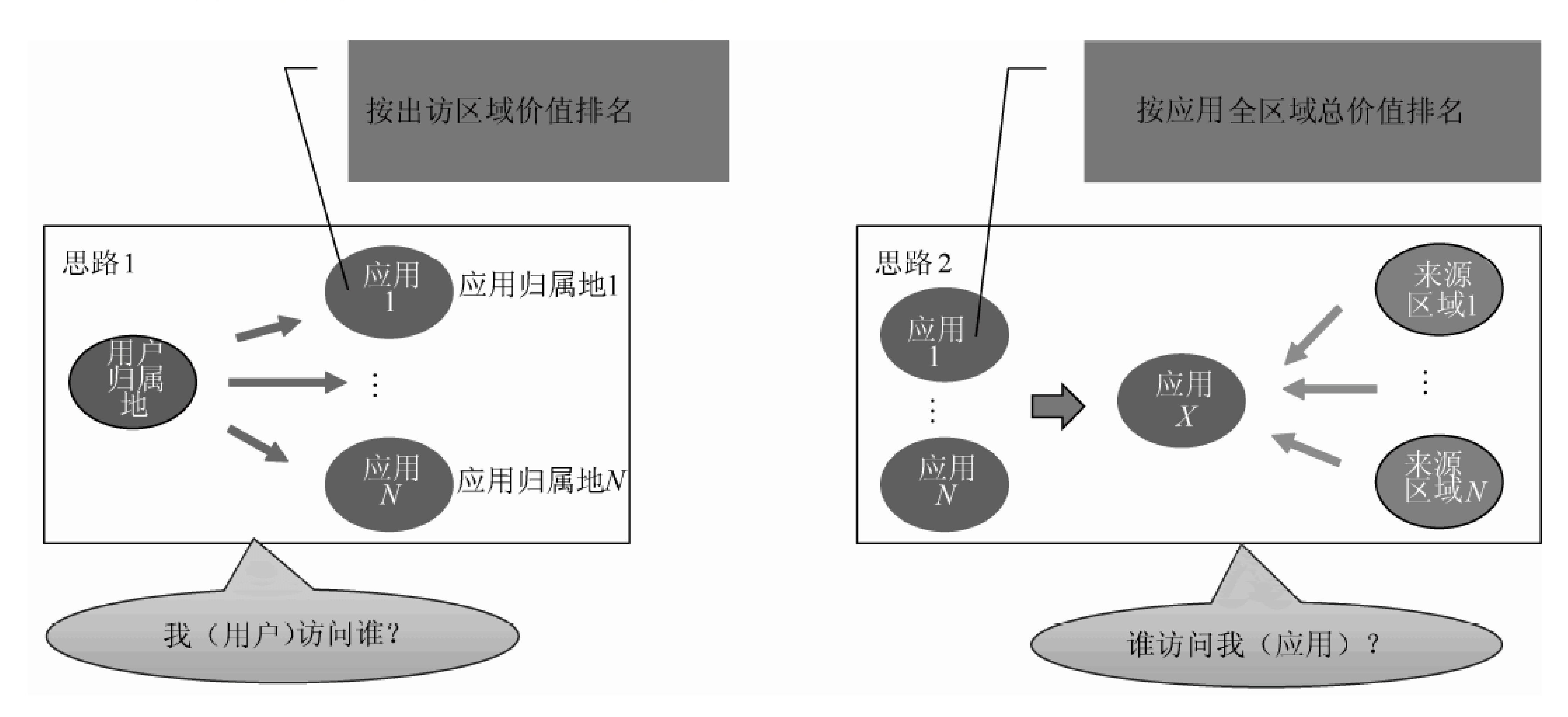


图 6-1-6　不同角度的价值排名

在图 6-1-6 中，思路 1 为从用户归属地到出访区域价值排名；在图 6-1-6 中，思路 2 为先从全区域应用价值排名，然后再以应用的访问来源区域价值排名。

需要注意的是，应用价值是流量带来的热度价值/总流量和区域用户带来的收入/用户带来的总收入的归一化结果。

以上方案的各个参与方均能够获得价值提升。移动用户可以获得更快的应用访问速度和更好的应用使用体验。应用提供商可以增强用户黏性，扩大用户规模。电信运营商则能够增加 IDC 业务收入。

3. 实施思路、方法及关键点

上述方案中思路 1 为从用户归属地到出访区域价值排名，CDN 节点设置方法如图 6-1-7 所示。

上述方案中思路 2 为先从全区域应用价值排名，再按应用访问来源区域价值排名，CDN 节点设置方法如图 6-1-8 所示。

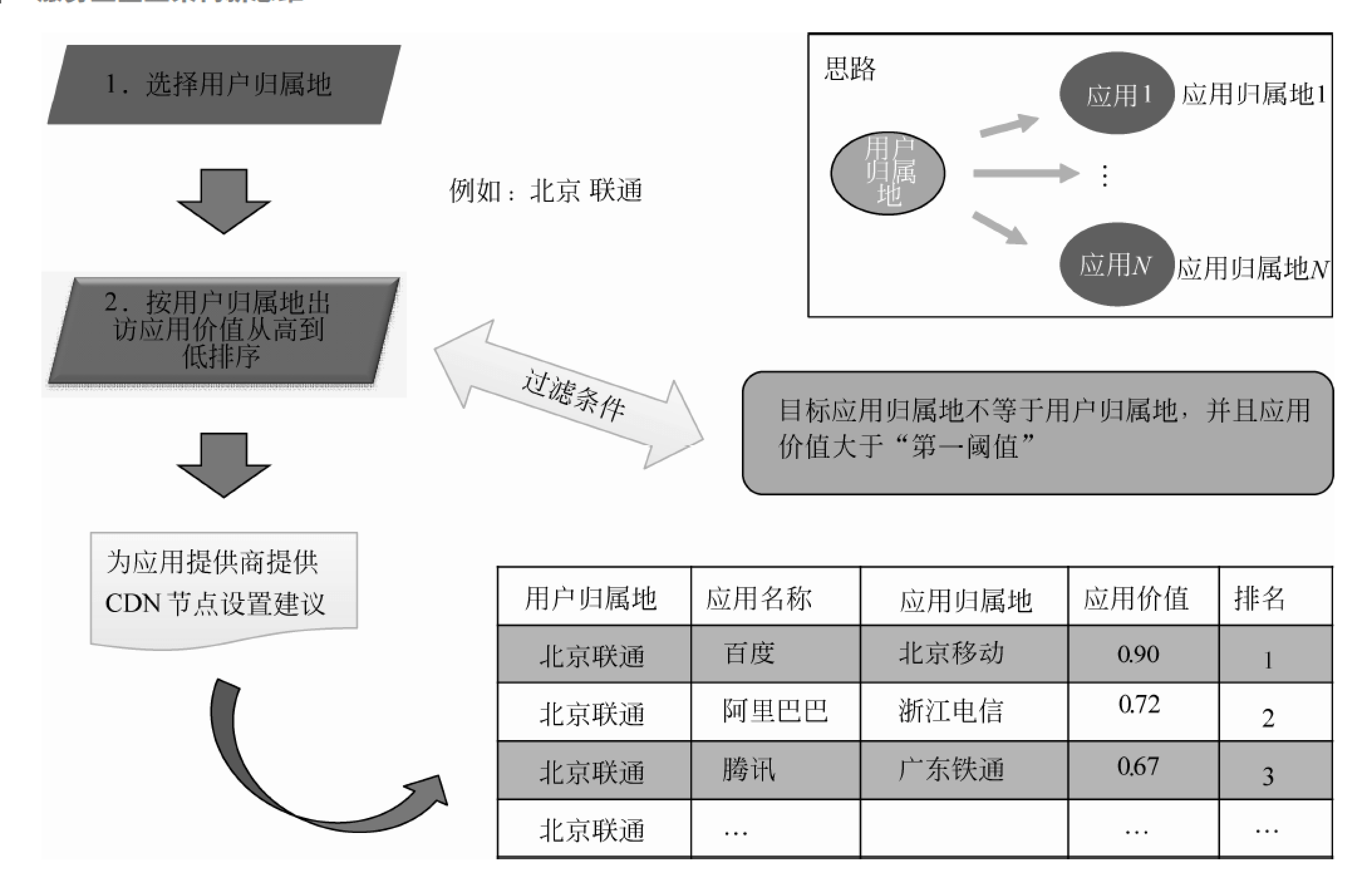


图 6-1-7　CDN 节点设置方法 1：从用户归属地到出访区域价值排名

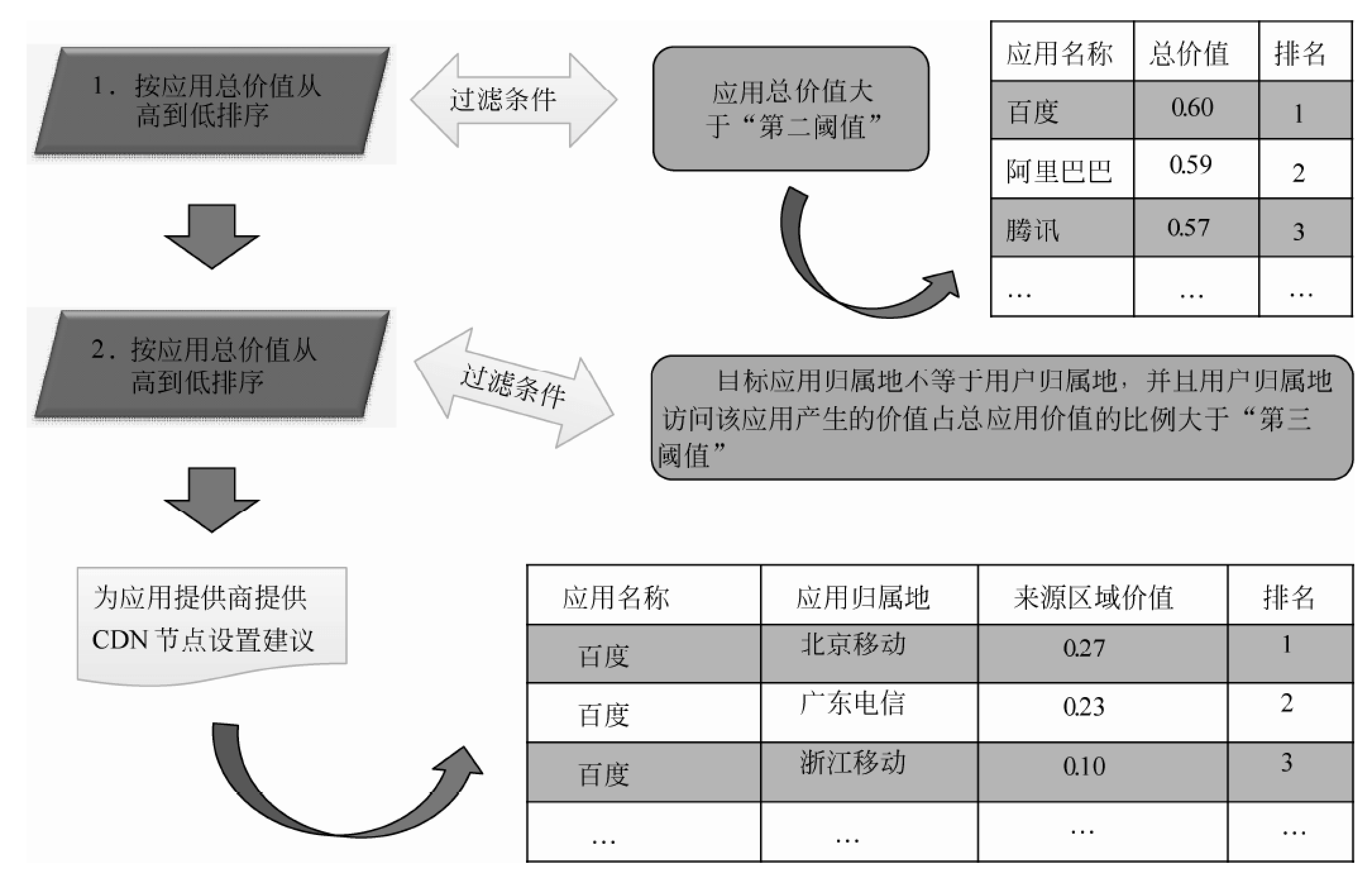


图 6-1-8　CDN 节点设置方法 2：先从全区域应用价值排名，再以应用访问来源区域价值排名

说明：在图 6-1-7 和图 6-1-8 所述的实施方案中，设置“第一阈值”、“第二阈值”和“第三阈值”，目的是为了过滤掉不符合要求的项目。

仅仅依靠移动用户上网记录大数据并不能计算出用户价值、应用价值以及网络访问路径，因此可以说移动用户上网记录是不完整的。

第一，移动用户上网记录中的关于应用的信息主要为网页 URL，并没有非网页应用的信息，因此要得到 IP 地址对应的应用信息，必须依靠与应用提供商合作，获取 IP 地址与应用映射关系数据，数据越丰富、越及时、越准确，就越能够正确地做出 CDN 节点设置决策。

第二，关于 IP 地址与区域映射关系，由于竞争关系和安全考虑，电信运营商往往难以获取其他电信运营商 IP 地址与区域映射关系的准确数据，这也是影响网络路径计算非常重要的因素。

因此，要想运用移动用户上网记录大数据解决 CDN 节点设置问题，关键是要不断改善数据的完整性。

6.2 大数据在金融行业的应用

金融的本质是信用，其作用是全社会资源配置，其管理的难点是风险，应当引全社会资源之水，灌溉资金供需之田，收获效率提升与风险可控之果。

金融行业具有银行、保险公司、共同基金等金融机构，经营存贷款、证券、基金、保险、债券等业务。金融机构通过货币经营，满足了货币供需参与方需求，实现了货币资金的重新配置，提高社会经济运行效率。

在金融交易中，对于货币或者实物的提供方，可以通过提供资金而获得利息、租金等报酬；对于货币或者实物的使用方，应当为资金提供方缴纳利息、租金等，对于提供各种增值服务的金融机构，则通过提供金融服务而获得经营收入。

金融的本质是信用。金融企业一方面需要管理好客户的信用，降低运营风险，另一方面，则需要通过多种渠道融资，降低资金成本和流动性风险。不同于提供产品和服务的生产型企业，金融企业的主要任务是管理金融风险。如果金融企业对收益一方的信用判断失

误，就会对金融企业造成经济损失。可见，信用管理对于金融企业至关重要。

在移动互联网和大数据时代，信息化已经逐步占据人类生产与生活的各个角落，企业行为和个人行为都被记录下来。由于各种信息日益公开和透明，企业属性和行为以及个人的属性和行为数据，成为金融机构完成信用评估的重要数据基础。

快速、准确地完成信用评估，可以帮助金融机构降低生产经营风险，提高市场竞争力，也可以帮助需要资金的企业或个人更快地获得所需的资金。可见，提升信用评估能力可以实现金融机构及其客户之间的双赢。

要完成信用评估，需要收集资金使用方的各种数据。对于银行贷款业务，需要收集客户的历史交易数据、工资水平、受雇企业的性质、学历高低等；对于汽车保险业务，需要收集客户的历史违章数据；对于生命保险业务，需要收集客户的健康数据，等等。不同的金融业务对于客户的关注点不同，通过基于对相关数据的收集和分析，可以达到信用评估的目的。当然，并不是所有的信用评价都可以直接量化的，有些信用评价指标还需要采用定性和定量相结合的方式计算出来。

为了提高信用评估的准确性和及时性，企业应当借助大数据，逐步减少定性因素，增加定量因素。

在未来的移动互联网时代，发展趋势是社会化分工更加细致，这是社会发展的必然要求：信息技术和互联网为新型的社会分工提供了工具和手段，越来越多的小微企业成为社会发展的新动力。因此，对于小微企业的信用评估也变得非常重要，笔者以阿里金融为案例，分析企业如何利用大数据完成小微企业的信用评估。

由于信用评估对象具有不同的属性和行为特点，因此对于个人、大中型企业、小微企业信用评估时，关注点是不同的。本章从信用评估对象的类型角度，分别给出实现大中型企业、小微企业以及个人信用评估的思路与方法。

6.2.1　应用场景 1：大中型企业信用评估新思路

从年销售收入和资产总额的角度看，年销售收入和资产总额大于 5000 万元的称为大中型企业。大中型企业有着与小微型企业不一样的特点。大中型企业的项目规模大，项目周期长、风险高，因此，对于大中型企业的信用评估也有着与小微型企业不一样的特点。

下面以两个国有大型金融企业对于大中型企业的信用评价方法为例，说明大中型企业信用评估的一般方法，然后再分析大数据时代，企业如何改进对于大中型企业的信用评估

方法。

1. 金融机构对企业的传统信用评估方法

首先以某大型国有银行的信用评估模型为例。某大型国有银行对于大中型企业信用评估主要关注被评企业的领导者素质、经济实力、资金结构、经营效益、信誉情况、发展前景等几个方面。

企业领导者对于企业的生产经营起着至关重要的作用，领导者素质包括领导者的教育经历、工作经历、工作能力、工作业绩、职业操守等方面；

企业经济实力包括实有净资产、有形长期资产、人均实用净资产几个方面，企业经济实力指标体现了企业的资产风险；

资金结构包括资产负债率、速动比率、流动比率、经营活动现金净流量几个方面，资金结构指标体现了企业的财务风险；

经济效益包括总资产净利率、销售利润率、利息保障倍数、应收账款票据周期次数，经济效益指标体现了企业的经营能力与经营风险；

信誉状况包括贷款质量、贷款付息、存贷款占比，信誉状况指标体现了企业的资金风险；

发展前景包括近三年利润情况、销售增长率、资本增值率、行业发展状况、市场预期状况、主要产品寿命、销售渠道，发展前景指标体现了企业经营风险。

企业信用评级体系采用百分制计分，分为 AAA、AA、A、BBB、BB、B、F 等级，从高到低进行排序，标识了企业所处的信用等级。AAA 级最高，代表着企业有很强的市场竞争力和很好的发展前景，企业流动性很好，管理水平很高，并且有很强的偿债能力；F 级最低，表示企业不符合国家环保、产业、信贷等有关政策，属于可疑或者损失类企业。

企业信用等级有效期为一年，如果在一年之内，企业经营状况发生重大变化，例如重大建设项目、重大法律诉讼、重大人事调整等，那么需要重新评级。

企业信用评级采用定性和定量相结合的方法，主要从市场竞争力、资产流动性、管理水平等几个方面评定，评级指标分类如下。

（1）市场竞争力：经营环境、质量管理体系、市场拓展和销售渠道等；

（2）资产流动性：流动比率、速动比率、应收账款周转率等；

（3）管理水平：主要管理人员素质和经验、资产报酬率等；

（4）其他方面：资产负债率、行业发展前景、重大事项分析结果等。

2. 大数据时代企业信用评估的新思路

通过对传统信用评估方法的分析可以看出，金融机构对于大中型企业的信用评估更多地从企业财务状况、经营状况、发展前景等几个方面来进行评估，更多地依赖信用评估人员的分析报告这样的定性评估方式，然后再用评分的方式进行量化，信用评估人员的知识能力和经验水平对信用评估结果起到非常重要的影响，信用评估结果存在很多主观因素。

从金融机构对大中型企业信用评估的方法可以看出，金融机构利用大数据实现信用评估是有很大难度的。金融机构可以在传统信用评估的基础上，逐步引入大数据，降低因评估人员的知识和经验形成的主观偏差，逐步加大信用评估模型中“定量”指标的比例，让信用评估结果更加准确，更具有决策参考性。

6.2.2 应用场景 2：小微企业信用评估新思路

小微企业是小型企业、微型企业、家庭作坊式企业、个体工商户的统称。大中型企业的特点是数量少、单个项目的信用风险大，因此对于信息系统的依赖性小，信用评价的难点为尽职调查阶段收集数据的可靠性以及评价人员的经验。

与大中型企业相比，小微企业数量多并且企业生命周期短，因此不可能像对待大中型企业那样进行尽职调查和信用评估。同时，小微企业在市场竞争环境中，不像大中型企业那样具有资源和市场优势，需要快速地完成信用评估并取得生产经营所需的资金。

1. 传统小微企业的信用评价方法

小微企业传统的信用评估分为财务因素和非财务因素两类指标。财务相关的指标包括偿债能力、经营能力、盈利能力等，非财务相关的指标包括企业领导人素质、企业素质、政策环境、合作关系等。

财务因素中的偿债能力体现了小微企业的抗风险能力，包括资产负债率、现金比率、主要资产、或有负债等；经营能力包括主营业务增长率、应收账款周转次数、纳税情况等；盈利能力包括净利润增长率、净资产收益率等。

非财务因素中的企业负责人应当是实际控制人，小微企业负责人的素质决定了小微企业的经营能力，包括企业负责人的个人品质、信用记录、从业年限、学历、健康状况等；企业的素质方面包括管理能力、管理团队、市场竞争力、公司成立年限、企业信用记录等；

政策环境对于小微企业也有很大的影响，包括行业集中度、行业政策、区域环境等；如果小微企业为仓储业，那么其生产经营场所、仓库面积利用率等因素对于信用评价也有很大的影响。

2. 大数据时代小微企业的信用评价方法

信息通信技术改变了人们工作和生活的方式，提高了社会效率和生活的便利性，小微企业的采购、销售等生产经营活动被记录下来。小微企业的行为痕迹对于小微企业的信用评价具有非常重要的作用。

由于小微企业具有数量多、融资频率高、融资需求额度小的特点，更适合通过借助系统快速实现对小微企业的信用评价。要完成快速的信用评价和放贷，金融企业势必要承担比传统信用评价方式更高的经营风险。

金融机构的工作难点在于金融风险的管理，而利率就是基于风险大小确定的，贷款预期风险越高，放贷利率越高，贷款预期风险越低，则放贷利率越低。统计学的大数定律理论说明：当试验次数足够多时，事件出现的频率无穷接近于该事件发生的概率，这是偶然现象背后存在的必要规律。根据大数定律理论，可以预见小微企业的平均贷款风险趋于预期贷款风险，因此可以利用小微企业的总体预期贷款损失率来代替每一笔小微企业贷款预期损失率，这样可以降低利率计算的难度，提升对小微企业的放贷效率，金融机构可以争取到更多的小微企业客户。

与小微企业相对应的是小额贷款。小额贷款具有期限短、额度小、随借随还的特点，因此更需要金融机构快速做出贷款决策。

放贷可以分为贷前、贷中和贷后三个阶段，要完成对小微企业的放贷，就需要快速完成对小微企业的信用评估，确定授信额度，并且通过对放贷后小微企业的生产经营行为进行实时监控和风险预警，尽早发现和规避风险，对于确认为具有金融欺诈行为的企业，应当采取严厉的惩罚措施。

在贷前阶段，主要任务是完成客户的初步授信工作。小微企业在电子商务平台上积累的交易记录是确定授信额度的主要参考内容，交易记录中具有小微企业的采购、物流、库存以及销售数据，可以反映小微企业的生产经营能力和财务能力。B2C 模式中客户对于企业产品和服务的评价、B2B 模式中供应商和合作伙伴对于小微企业的评级，也是信用评估的重要数据来源。此外，还可以从金融管理机构（比如中国人民银行）获取小微企业的信用记录。

在贷中阶段，主要完成对小微企业的审查工作，审查目标是确定小微企业经营者的诚信度。由于小微企业数量众多，可以采用分析远程视频采访录像的方法，测试小微企业经营者贷款意图是否存在撒谎行为。

在贷后阶段，主要目标是降低小微企业的本金和利息偿还风险。企业可以通过监控小微企业资金运用情况，掌握小微企业的贷款是否按照事先的计划从事生产经营活动，比如是否将贷款用于广告投放并因广告投放而增加了交易数，是否将资金用于采购销售品等。如果发现小微企业贷款后并没有出现采购、营销等行为，并且销售量也没有因新的资金注入而发生变化，那么就需要进行风险提醒和预警。对于未按合同约定逾期还款的，则需要按约定支付罚息，对于逾期一定期限未还款的，则需要将该小微企业放入黑名单并进行全网通缉，进行更加严厉的制裁，让该小微企业为不诚信行为付出很高的代价。

阿里巴巴是互联网企业为小微企业提供小额贷款服务的典范，可以为处于弱势地位的小微企业提供传统金融渠道无法提供的小额贷款服务。阿里巴巴具有像淘宝和天猫这样面向大众的 B2C 电子商务平台，有面向供应商和批发商的 B2B 电子商务平台 1688 等，这些电子商务平台中记录的市场推广、交易、评价等数据成为对小微企业授信的“信息流”维度数据源；支付宝可以记录小微企业的现金流，成为“资金流”维度的数据源；“菜鸟网络”等物流平台上记录了小微企业的采购、库存等数据，成为“物流”维度的数据源。阿里小微金融信用评估数据体系如图 6-2-1 所示。

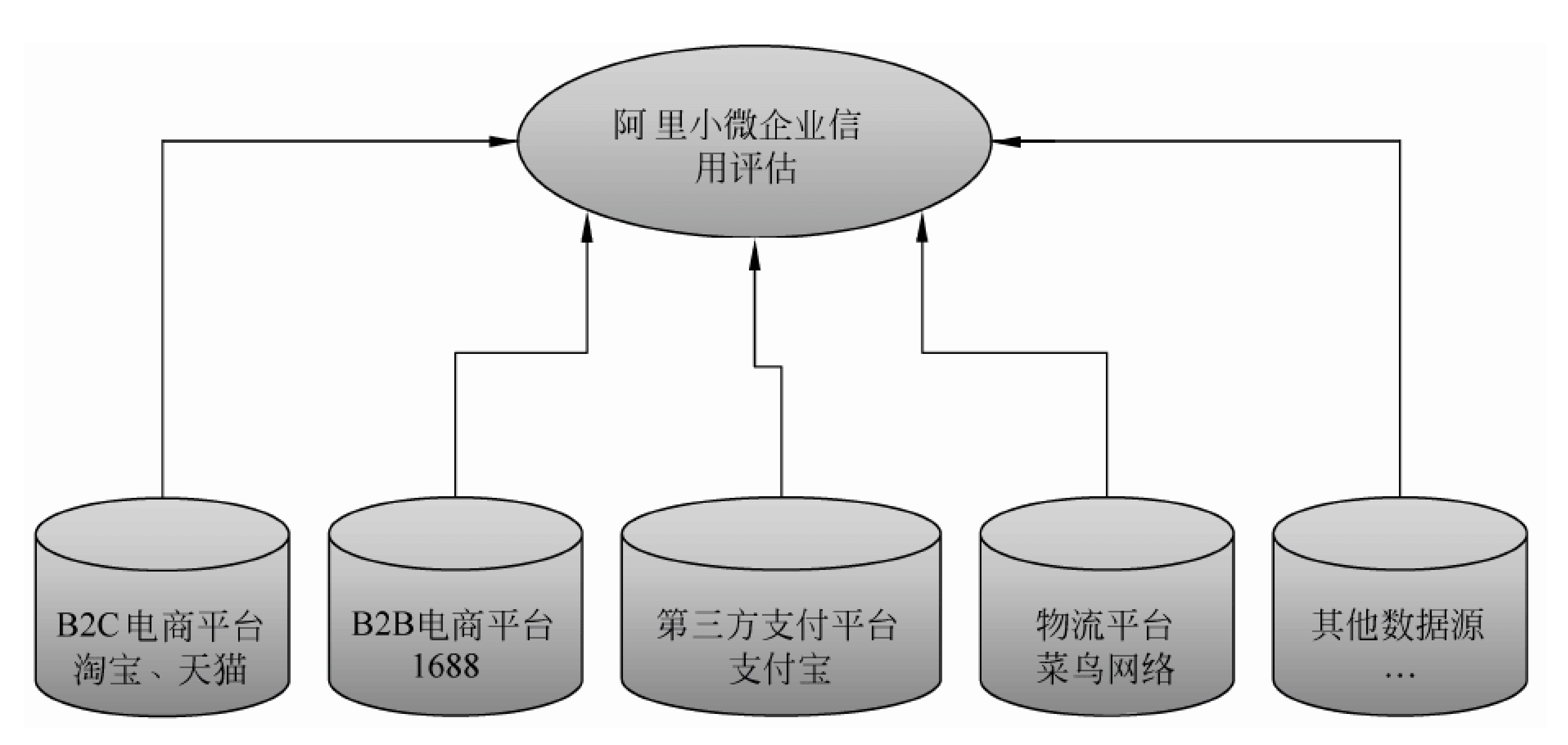


图 6-2-1　阿里小微金融信用评估数据体系

丰富完整的电子商务数据源成为阿里巴巴对小微企业授信的数据基础和关键，成为阿里巴巴在小微企业信用评估领域独特的竞争力，来自外部的其他数据源对于信用评估也起到重要的辅助作用。此外，采用视频采访、社会关系调查等获取的非结构化数据和软数据等提高了信用评估的准确性，成为信用评估的重要补充。

6.2.3 应用场景 3：个人信用评估新思路

1. 个人信用评估的典范：FICO 评分系统

FICO 评分系统由成立于 20 世纪 50 年代的费埃哲（Fair Isaac）公司发明，该公司名字为创始人 Bill Fair 和 Earl Isaac 两人姓名的首字母。随着计算机技术的发展和应用，FICO 评分系统得到了快速而广泛的应用，并逐渐成为美国征信业事实上的国家标准。

1970 年，费埃哲公司开始向银行出售信用评分。1989 年，开始开发 FICO（费埃哲）分数。FICO 面向个人征信，分数范围为 300 分到 850 分。其中，680 分以上为信用卓越，620 分以下则需要增加担保或者拒绝贷款。FICO 可以快速、客观地量度个人风险。信用信息越早，对于信用评估的影响越小。美国的三大征信公司：益百利（Experian）、艾克发（Equifax）和环联（Trans union）的信用评估模型都是以 FICO 为基础的，评估模型和评估结果均差别不大。

FICO 评分系统中包含了完整的个人信用信息和多年（7～10 年）的个人信用记录。个人信用相关的信息包括来自商业部门和社会公共事业部门的记录。商业部门包括银行、保险、证券等，比如银行信用记录、保险信用记录、证券信用记录等；社会公共事业部门包括公安、法院、税务等，比如违法犯罪记录、法律诉讼记录、所得税缴费记录等。

FICO 关注的因素分为 5 类，即客户信用偿还历史、信用账户数、使用信用年限、正在使用的信用类型以及新开立的信用账户。下面分别进行说明：

1）客户信用偿还历史

是 5 个因素中最重要的因素，在整个信用评分中占比大约为 35%。具体包括：

- 各种信用账户的还款记录，包括信用卡、分期偿还贷款、抵押贷款等；
- 公开记录及支票贷款记录，包括破产记录、法律诉讼事件等；
- 预期偿还情况，包括逾期天数、未偿还金额、逾期还款次数、逾期发生距离现在的时间长度等。

2）信用账户数

主要反映客户的整体还款能力，比客户信用偿还历史的重要性要低，在整个信用评分中占比大约为 30%。

3）使用信用的年限

体现了信用账户的账龄，在整个信用评分中占比大约为 10%。

4）正在使用的信用类型

主要是针对多个账户混合使用的情况，对于客户来说，可能会具有多个不同类型的账户，比如信用卡账户、零售账户、分期付款账户、金融公司账户、抵押贷款账户等，不同的账户类型其风险系数是不同的，因此要区别对待。正在使用的信用类型在整个信用评分中大约占 10%的比例。

5）新开立信用账户数

体现了客户信用风险的程度，如果客户在短时间内开立了多个信用账户，那么该客户的资金偿还风险一定会高一些。新开立信用账户数在整个信用评分中大约占 10%的比例。

此外，为了尊重个人隐私，种族、肤色、宗教、性别、婚姻状况等个人基本信息在 FICO 评分系统中不参与评分，而工资、职业、头衔、雇主、受雇时间、受雇历史等工作相关信息仅仅作为评分的参考项，同样不直接参与评分。

2．我国个人信用评估案例：WeCash

FICO 评分系统是根据美国国情而发展起来的，许多评价方法在我国并不适用，我们一方面需要参考 FICO 的信用评价方法，同时也需要政府部门构建信用评估方法和体系。商业机构和社会公共事业部门需要对外开放信用评估所需的数据，借助科学、客观、有效的评估指标和评估方法，逐步完善金融体系，帮助企业更加精确地估计消费信贷风险，提升工作效率。

2014 年，国内首家大数据信用评估公司 Wecash（闪银）获得 IDG 4000 万元的 A 轮投资，公司估值两个亿，拉开了国内采用大数据实现企业信用评估的序幕。

在信用价值领域，我国境内所谓的信用卡具有申办流程和所需资料复杂，办理期限冗长，银行信用价值在实际生活中的应用非常受限，而 Wecash 可以通过大数据分析和机器学习，对传统银行的信用评估模型进行精简，无须提供材料，大部分数据基础是个人的 SNS 数据、互联网搜索数据等行为数据，可以将整个评估流程控制在 20 分钟以内。此外，Wecash 还可以利用互联网行为转化为“互联网信用”，从而拓展了应用场景，例如可以应用于公

司招聘等。

6.3 大数据在互联网行业的应用

互联网强调平等、协作、去中心化，通过搜索、社交、购物等互联网应用沉淀下来的海量数据，成为推动社会创新发展的催化剂。

互联网始于 1969 年的美国，又称因特网，以“开放、合作、创新”为特征的互联网经过近半个世纪的发展演进，发生了翻天覆地的变化，大大改变了人类的工作与生活。

互联网经济也称为眼球经济、无摩擦经济，可见互联网公司必须积攒人气才行。因为其进入门槛比较低，竞争激烈，互联网公司必须借助丰富的内容和优质的服务来吸引并留住客户，增强客户黏性是互联网公司的第一要务。

在互联网上，人们不仅可以浏览与分享信息、沟通交流，同时也可以进行交易。与传统商业模式不一样，基于互联网的交易直接打通买卖双方，减少了渠道分销等中间环节，因此又称为无摩擦经济。

互联网公司为客户提供各种产品和服务的同时，留下了大量的接触“痕迹”，比如浏览、搜索、登录、退出、下单、投诉、咨询、建议等，这些行为轨迹可以反映客户特征，让互联网公司更好地把握客户需求，推送符合客户需求的产品和服务，提升产品销售能力。

大数据与互联网的关系最为紧密。大数据在互联网领域的典型应用就是搜索，搜索是人们通过互联网获取信息的入口，搜索服务的基础就是 Web 内容大数据，Web 内容是半结构化的，同时由于互联网人人都是内容的创造者，因此 Web 内容产生的速率也是非常快的。以谷歌、雅虎为代表的互联网公司解决了 Web 数据海量存取问题，成为大数据技术发展的先行者。

本章主要分析大数据在社交网络和电子商务领域的应用。

6.3.1 应用场景 1：大数据在社交网络领域的应用

1989 年，万维网之父蒂姆·伯纳斯·李（Tim Berners-Lee）发明了 World Wide Web，即当前互联网应用广泛的 WWW（3W）。伯纳斯·李认为 Web 的最终目标帮助人们实现像

Web 一样的存在方式，直至 20 年后的今天，WWW 确如伯纳斯 • 李希望的那样，WWW 就像蜘蛛网一样，渗透到人们工作与生活的各个方面。

Web 发展的三个阶段，业界将其定义为 Web1.0、Web2.0 和 Web3.0。

Web1.0 阶段以门户网站为代表，信息以类似于广播电台的单向方式传播，用户通常是借助 Web 门户获取信息的，典型应用有 BBS、新闻网站。

Web2.0 体现了互联网人人参与的思想，每个网民既是 Web 内容的消费者，同时又是 Web 内容的提供者，典型应用有博客、Wiki、IM（即时消息）等，Web2.0 给予了广大草根网民参与其中并受到关注的机会，也大大丰富了互联网的内容。

Web3.0 以社交网络系统（Society Network System，SNS）为代表，更强调网民之间的沟通与协作，同时，Web3.0 中的应用提供商通过构建能力开放平台，使得网民都可以参加到 Web 应用创新之中，网民既可以是 Web 应用的需求提出方，同时也可以是 Web 应用的软件开发方，人人都可能通过努力具备一定的影响力而成为“明星”，也可以成为自己喜爱的明星的“粉丝”。Web3.0 时代的典型应用包括微博、微信等。

1. SNS 业务应用介绍

SNS 体现了 Web 对于人类社会需求的满足，人们通过 Web 应用满足不同的心理需求。SNS 不同于支持组织活动和业务流程的传统应用，其主要实现了人与人以及由人创建的内容之间的协同和共享。

SNS 体现了人类的社会性，一个人从出生到成长会形成各种各样的社会关系，比如家庭关系、同学关系、同事关系、战友关系、老乡关系、朋友关系等。

家庭关系网络：家庭是一个人出生和成长的起点，会因各种血缘关系伴随人的一生，包括父母、兄弟、姐妹、七大姑八大姨等。

同学关系网络：同学关系是在一个人接受思想道德教育和智力教育的过程中形成的，在接受教育过程的不同阶段会形成小学同学、中学同学、大学同学等关系。

老乡关系网络：老乡关系则是由于社会生活的流动性引起的，一个人可能会因为工作和生活需要而离开自己的家乡，在家乡之外如果能够遇到与自己具有类似口语、类似风俗习惯的人会倍感温暖亲切。

同事关系网络：同事关系是在个人工作过程中建立的，同事之间通常在所处行业和工作内容方面具有很大的相似性，因此成为猎头公司发现人才的好途径。

朋友关系网络：朋友关系建立在共同的兴趣爱好之上，所谓“物以类聚，人以群分”，

朋友关系更多的是具有接近的脾性，因此可以作为发现用户偏好的一种方式。微信的朋友圈就是基于朋友关系网络的一个典型移动互联网应用。

SNS 在满足社会沟通和协同方面分为多种类型，包括商务类、娱乐类、婚介类、娱乐类、综合类等。国外著名的 SNS 包括 Facebook、Twitter、Linkedin、GitHub、WhatsApp 等，国内著名的 SNS 包括微信、微博、人人网（原校内网）、朋友网、开心网、百合网、珍爱网等。

微博模式为“明星-粉丝”模式，微博用户不一定是一个真实存在的人，它可以是一个公司、想象中的人甚至是已经不在世的人，当明星发表言论后，粉丝通常会跟随，发表评论。微博是现代社会节奏加快，需要快速简短地表达自己的想法的一种体现，每条微博的总字数通常不超过 140 个字，这有些类似于《读者》杂志，不同于中长篇小说和著作，每篇文章只是阐明生活的某一方面的观点，篇幅都在几千字之内，读者可以像吃一顿快餐一样快速地完成阅读。

Linkedin 是面向商务人士的职业社交网站，Linkedin 基于个人所在工作单位的名称、职务、专业方向、地理位置等建立人与人之间的关联关系。

GitHub 则面向程序开发者，开发者可以借助 GitHub 来分享源代码。

在我国，工作与生活通常是不分的，因此微博、微信等社交网络应用通常是同时面向工作关系和生活关系的，在美国则不同，Linkedin 专注于商务关系，Twitter 则专注于生活关系。

2. 大数据技术与 SNS 应用

不同的社交网络应用的特点是不同的，微博类应用主要反映热点话题，因此要求大数据技术能够实现热词的提取和统计，利用自然语言处理技术来分析评论内容；面向商务人士的社交网络应用则需要按照工作职位、单位名称、专业方向、兴趣爱好等将人与人关联起来。

除了人们在 SNS 上形成的社会关系以及留下的沟通、评论等记录，用户的通信行为也是反映用户社会关系的重要数据基础。可以以用户打电话、发短信这样的通信行为为基础，形成用户之间的通信行为网络，通信行为网络中的每个“点”就是具备外呼行为的通信用户，两个用户之间形成的边就是用户之间的通信行为，“边”上包括两个用户的通信时间、通信地点、通信时长等，如果对于通信次数进行统计，那么通信次数多的两个用户的社会关系是紧密的，通过统计可以发现用户之间社会关系的强弱。

6.3.2 应用场景 2：大数据在电子商务领域的应用

电子商务是商务活动的电子化，由信息流、资金流和物流三大要素构成，主要包括商家对客户（Business to Customer，B2C）、商家对商家（Business to Business，B2B）两种模式。

大数据对电子商务的主要作用是发现用户行为，然后有针对性地为客户提供产品和服务。从客户角度看，客户需要经过商品发现、商品购买、服务获取三大阶段。

在客户的商品发现阶段，客户通过搜索、浏览方式来发现商品和对比商品，企业可以利用大数据技术提供热搜商品排行榜，对客户浏览的网页和时长进行统计，发现客户感兴趣的商品。企业可以结合用户特征，从购买类似商品的视角为客户提供商品推荐，比如“购买了该商品的用户还购买了 XX 商品”。

在客户的服务获取阶段，企业可以基于客户咨询、投诉、建议、评价等记录分析客户对于哪些商品感兴趣，对于某些商品的看法，辅助调整商品采购列表，为供应商提供商品改进建议，提高服务质量等。可以借助大数据技术，将客户经常提出的问题进行整理并形成知识库，提高客户服务的效率。

6.4 大数据与隐私保护

信息共享和数据开放既是把双刃剑，能否为造福人类关键要看我们的态度和行动，只有构建科学的组织、制度和流程，才能趋利避害，实现共赢。

6.4.1 科技进步的代价

随着科学技术在信息、通信、生物等领域的飞速发展，人们的工作与生活进入了快速、全面的“记忆”时代。

从记忆的方式看，“电脑”时代之前，信息通常由大脑或者纸张“记忆”下来，信息交换的方式通常是言语沟通或者印刷品传播，这个阶段信息传播的特点是传播范围小，传播速度慢。在当下的互联网时代，承载信息的方式是互联网，信息可以瞬间全球传播与共享。

先进工具的使用，就好比在沙滩上行走一样，会留下印记，而且这种印记让人难以察觉。例如：

- 交通管理部门记录了人们的出发地、目的地等信息；
- 酒店机构记录了人们的住宿地点、房间号、陪同人员等信息；
- 旅游公司记录了人们的旅行路线、地点等信息；
- 金融机构记录了人们的存款额度、理财产品、参保类型等信息；
- 医疗保健机构记录了人们的身高、体重、血压、所患疾病等信息；
- 互联网公司记录了人们的网页浏览、关键字搜索、社交网络、网络购物等信息；
- 公共管理部门记录了人们的水电煤等的使用时间、使用量等信息，等等。

不同企业的“猜测”能力是不一样的。亚马逊（Amazon）监控购买偏好，谷歌（Google）知道浏览习惯，推特（Twitter）知道人们所想，脸书（Facebook）不但知道人们所想，而且知道人们的社交关系，移动运营商知道人们和谁交谈并且谁在附近。商家通过信息采集与数据分析，确定营销与服务的时机、对象、内容等前提条件。商家对用户了解得越多，越能够影响决策：

- 通过分析搜索关键字，知道用户关注什么内容、大家都关注什么内容；
- 通过分析网页浏览次数、停留时间，知道用户关注什么商品，大家关注什么商品；
- 通过分析人们在因特网上搜索、浏览、咨询、下单、退货等行为，可以快速获取到电话号码、邮箱、所在位置、偏好等个人信息。

新的技术手段可以用于确定商业规则，但也可能会留下侵犯隐私的隐患。例如，保险公司车险费率采用汽车上安装定位装置的技术手段，根据驾驶员的驾驶情况确定车险费率，如果驾驶情况良好则费率低，否则费率就高，但是这种行为有可能会侵犯驾驶人员的个人隐私。地理位置信息可以用于确定物体的位置，也可能会触犯正在屋顶晒太阳的人的个人隐私。

在个人生活方面，当人们在看病或者体检时，身高、体重、病史、DNA 等身体生理特征信息立马被医疗机构获取。人们的生理特征和行为特征也能够被记录下来，比如手印、DNA、气味、视网膜、声音、手势、打字节奏等，这些信息可以用于确认身份的真实性。

随着移动智能终端和移动互联网应用的飞速发展，智能手机作为人们的贴身小秘书，已经从单一通话功能延伸到照相、录音、办公、导航、上网等多种功能，当在智能手机上安装软件时，软件提供商首先会提示用户接受一系列控制权限，比如电话本、通话记录、短信记录、照相、录音、个人信息、应用信息、位置信息等。

此外，为了提升数据质量，可以通过很多途径反算出个人信息，比如 Web 搜索、电影

评论、上网记录、通话记录等。比如，人们在某个移动通信基站下打电话，可以根据通话记录中的基站编号、基站经纬度、基站覆盖半径等，反算出通话的人当时所在的位置。

与科学技术落后的年代相比，人们的社会生活变得更加便捷、高效。但是，任何事物都具有两面性，种种便利背后的代价，就是人们的隐私可能受到侵犯，人们的安全可能受到威胁。社会对你的“记忆”越多，个人的隐私和安全受到威胁的可能性就越大。大数据是个双刃剑，就像菜刀一样，关键看在什么样的手里，是用于做菜还是用于伤害。

棱镜计划（PRISM），俗称“棱镜门”，是一项由美国国家安全局（NSA）实施的绝密监听计划，自 2007 年小布什时期起开始实施，包括微软、雅虎、谷歌、苹果等在内的 9 家国际网络巨头都参与其中。NSA 可以直接进入美国网际网络公司的中心服务器，实施情报收集和数据挖掘。

“棱镜门”成为震惊国家安全领域的大事件，也说明了信息安全不仅仅是个人的事情，它将安全问题提升到国家层面。

6.4.2 人们应该做些什么

现实不能改变，能够改变的只能是自己的态度和行动。是否能够创建一个良好的隐私保护和安全管理环境，还有赖于个人、企业和政府的共同努力。

对于个人，人们应当了解信息公开对于个人可能造成的伤害。比如，对于未知来源的网页或者邮件，不要轻易打开，以免电脑中潜入木马程序，可以安装安全管理软件，实时监控入侵行为并进行及时清理。

对于提供软件服务的商家，需要在用户安装软件时，提示获取用户哪些信息，让用户可以自行选择是否安装，通过协议承诺和行动打消用户心中的顾虑；要在企业内部建立隐私保护制度和流程，防止内部员工盗取涉及个人隐私和安全的信息。

技术的进步总是超前于法律制定，对于政府管理部门，应当制定并细化保护隐私和信息安全的法律法规，对于触犯法律的行为要给予严厉的惩罚。

6.4.3 寻求法律保护

美国是隐私保护法制定的先行者。美国政府将海量的数据定性为有价值的国家资本，认为应对公众开放数据而不是禁锢在政府的体制内。联邦政府开放信息后，普通的公民都

可以享用政府提供的信息。

《信息自由法》的草案由摩斯先生在 1955 年提出，经过十几年的曲折历程，在 1967 年开始生效，直至 1974 年的《信息自由法修正案》才正式成为法律。

同样是 1974 年，美国国会通过了《隐私法》。《隐私法》保护的主体是存储在政府机关内部的“个人信息记录”，如个人的教育经历、工作履历、经济活动、犯罪历史等。

关于美国法律在信息自由、数据开放以及隐私保护方面的发展历程，《大数据》作者涂子沛先生有很多描述。

我国宪法中也明确了对于隐私的保护。

宪法第三十八条规定：“中华人民共和国公民的人格尊严不受侵犯。”人格尊严是人格权的重要内容，是人格利益的集中体现。宪法对人格尊严的规定，为我国日后完善隐私权制度提供了宪法依据。

宪法第三十九条规定：中华人民共和国公民的住宅不受侵犯。禁止非法搜查或者非法侵入公民的住宅。”这是宪法对公民私生活免受干扰的规定。

宪法第四十条规定：中华人民共和国公民的通信自由和通信秘密受法律保护，除因国家安全或者追究刑事犯罪的需要，由公安机关或检察机关依照法律规定的程序对通信进行检查外，任何组织或者个人不得以任何理由侵犯公民的通信自由和通信秘密。

自由和责任是一对孪生兄弟，信息共享和数据开放既是人类的福音，同时也可能对人类造成伤害，这是大千世界的不二法则，问题的关键在于人们管理数据的努力。魔高一尺，道高一丈，相信人类在不断的矛盾斗争中，必将能够趋利避害，战胜因信息与数据开放带来的种种困扰。

6.5　大数据相关热点话题

云计算为大数据提供弹性的基础设施，移动互联网、物联网、电子商务既是大数据的提供者，又是大数据服务的消费者。

6.5.1　概述

云计算、移动互联网、物联网、电子商务等是与大数据并驾齐驱的社会热点，这些社

会热点从不同侧面反映了商业、技术与社会的发展趋势，下面就逐个分析这些热点产生的背景、内涵以及其与大数据的关系。

6.5.2 云计算

1. 云计算的产生和发展

水能、风能、蒸汽能、煤能等都可以转化为机械能、电能，帮助人类提高生产和生活的效率，改善人类的生活质量。尤其是电，可以做的事情更多，人们日常生活中用的电车、电冰箱、电视、电灯、电话、电脑、洗衣机等，都缺不了电。正如人类生存离不开水和阳光一样，如果没有了电，社会生活的秩序将会遭到严重破坏。

电的发展经历了一个从分散到集中的过程。早期由于技术的限制，人们只能采用小的电厂发电。然而，这种方式不如集中建设电厂、集中供电更能节约成本，提高电的利用率，因此供电系统转变为集中建设、集中维护、统筹供电的运作模式。

从电力的发展模式可以得出一个结论：随着技术的进步，必然会用集中化建设和运营的模式代替传统资源分散的落后模式，发挥规模经济的优势，这是事物发展的必然。

信息技术的发展历程与电力的发展历程类似，通过软件将分散的资源集中起来，实现资源的统一调配，以最佳成本效益的方式提供 IT 服务。

简单回顾一下 IT 发展的历程：大约在 1995 年之前，软件基本上停留在小范围内使用，软件功能也比较简单，这个时候大多数企业和个人采用购买软件产品的方式来满足自身的业务需求；大约在 1996 以后的 10 年间，这种方式仍旧在继续，但是软件开始以服务形式对外销售。软件以服务形式销售的模式，称为软件即服务（SaaS），国外以 Salesforce 公司最为典型，国内以金蝶公司最为典型。

其实，像谷歌、百度等互联网服务提供商一直以来就是提供软件服务的，只不过在云计算兴起之前没有 SaaS 的叫法。Salesforce、金蝶等主要是将传统的套装软件以服务的形式放到互联网上销售。

除了在软件层面的 IT 服务，在计算、存储、网络层面也开始以 IT 服务的形式对外销售。在计算服务方面，曾经出现了分布式计算、并行计算、网格计算等研究和应用方向，在存储服务方面，也出现了存储虚拟化等研究和应用方向。

从服务范围角度看，云计算分为公有云和私有云。公有云是面向社会大众提供云服务

的，主要面向中小型企业和个人。私有云是面向单个组织内部提供云服务的，包括政府机关、事业单位以及大中型企业内部搭建的云平台。

从系统架构的角度，云计算分为软件层云服务（即 SaaS）、平台层云服务（即 PaaS）和基础设施层云服务（即 IaaS）。

从 IT 职能分工的角度，云计算分为计算云服务、存储云服务（也称为云存储）、网络云服务（Network as a Service，NaaS）、桌面云服务（Desktop as a Service，DaaS）等，分别提供计算、存储、传输等方面的云服务。

云计算还有很多其他划分方法，此处不再一一说明。

云计算的优势主要包括资源按需分配、后台能力扩展性好、高性能以及成本节约几个方面。

- 资源按需分配：使用云服务的用户无须关心后台资源如何分配和调度，只需提出能力需求即可，对用户来说后台完全是一个黑盒子；
- 后台能力扩展性好：当发现 IT 能力不足时，提供云服务的后台可以动态增加资源，满足 IT 能力需求；
- 高性能：采用云计算后，由于实现了分散资源的共享，IT 能力不再由单一的节点来支撑，因此可以提供更好的系统性能；
- 成本节约：云计算技术可以将组织内部现有设备充分利用起来，并实现资源的共享，通过软件算法保证系统的可靠性、可用性、可伸缩性、高性能以及安全性，从而降低了组织的采购成本。

2. 云计算与大数据

云计算和大数据就像是汽车发动机和汽油的关系，大数据提供动力所需的能量基础，而云计算则基于将“能量”转换为“动力”，使得汽车能够动起来。如果没有云计算，则大数据这个能源宝藏就得不到有效开发和利用，如果没有大数据，那么云计算则英雄无用武之地。大数据的首要特征就是数据规模大，这更加凸显了云计算的价值和作用。

云计算技术将数据随机存放到分布式的存储系统节点中，而不是以传统方式存放到预先设定的节点上，因此云服务的用户并不清楚云服务提供节点的位置，这种不透明性增加了隐私保护和提升数据安全管理能力的难度。

阿帕奇开源项目 Hadoop 中，HDFS 属于大数据的承载体，而 MapReduce 则是云计算的化身，HDFS 将大文件“微分”后存入集群中，当需要统计时，通过 MapReduce 实现对

“微分”数据的“积分”，经过抽取、排序和聚合的过程，输出计算结果。

6.5.3 移动互联网

近年来，移动终端和无线通信网络技术的快速发展使得人类社会进入了移动互联网时代。与个人电脑相比，移动终端具有随身性，无线宽带化使得人们可以借助移动终端实现桌面终端完成的一切事情，并且应用访问不受时空限制，这大大地方便了人们的工作和生活，成为移动互联网快速发展的前提。

移动互联网除了为用户带来使用上的便捷性之外，以 Apple Store 为代表的应用商店模式催生了更加丰富的应用，促进了信息通信产业的进一步发展，形成了更加专业化的社会分工，人类社会进入价值网络时代。

移动互联网加速了数据产生的速度。人们可以通过手机随时随地分享照片、上网聊天、发表看法，记录生活中的点点滴滴，人人都是自己生活的“记者”，同时人人又是内容的“消费者”。

移动互联网能够更加准确地掌握用户行为。与桌面互联网不同，用户的位置信息和行为信息，会实时地记录下来。因此，移动互联网能够知道用户在哪里、正在做什么以及下一步可能做什么。

移动互联网使得企业更能把握商机。商家可以借助移动互联网对于用户位置和行为的掌握，有的放矢地提供产品和服务，提升销售能力和服务能力。移动互联网也使得人们可以随时随地处理邮件、开多媒体会议等。

移动互联网使得政府行政部门更能够把握群体行为。政府部门也能够更加准确及时地预测群体行为，提前做好预案，增强公共管理能力。比如，交通部门可以节假日交通工具需求预测，提前准备交通工具。

移动互联网使得国家公共安全部门的破案能力更强。公共安全管理部门也可以通过分析用户特征和行为，发现犯罪分子的犯罪特征，预测犯罪行为并提前采取行动。

大数据是移动互联网发挥能力的基础。移动用户使用移动互联网应用时形成的大数据是移动互联网在商业、公共事业管理、公共安全管理等领域创造价值的前提和基础，移动互联网的发展离不开大数据技术的发展。

基于实时位置的移动互联网应用对大数据技术提出了新要求。阿帕奇项目（Apache）的 Storm 和 Spark 开源框架主要解决大数据实时流式计算问题。Storm 可以实时统计移动互

联网产生的数据，比如热词统计、用户画像更新等。

6.5.4 物联网

顾名思义，物联网（Internet of Things）是实现物理对象互联互通的网络。《大话物联网》作者郎为民老师认为："如果说因特网让全世界变成了一个村，那么物联网就让这个村变成了一个人；如果因特网连接的是虚拟信息空间，那么物联网连接的就是现实物理世界；如果说因特网是人的大脑，那物联网就是人的四肢"。

通过物联网与因特网的对比，可以清晰地看到物联网的重点是"物"之间的连接，而因特网的重点是"信息"之间的连接。"物"是人的肉体之外的东西（Things），而"信息"则是对客观世界中各种事物的运动状态和变化的反映。

技术是物与物连接的基础。物联网相关的技术包括射频识别技术（RFID）、传感器（Sensor）技术、纳米技术、智能嵌入式技术等。

物联网技术已经广泛应用于交通、物流、建筑等各行各业。在交通行业，利用 RFID/NFC 技术实现快捷支付；在仓储物流行业，利用 RFID 技术可以快速实现商品的入库、出库等操作；在商品零售行业，利用 RFID 技术可以快速实现商品的盘点、收银等操作；在建筑行业，利用 Wi-Fi、蓝牙（Bluetooth）等无线通信技术对家用电器进行远程控制，实现智能化家居生活等。

物联网技术还可以完成自然环境的监测。各种传感器网络相当于人的四肢，可以从自然环境中采集温度、湿度、风力、气压等数据。利用传感器采集的数据，可以辅助完成农业种植、工业控制、气象预报等工作。比如，在农业种植方面，可以根据传感器监测的湿度数据确定灌溉用水的量；在工业控制方面，可以根据传感器收集的工业设备温度数据调整车间空调温度；在气象预报工作中，更需要利用传感器采集的风力、风向、温度、湿度等数据进行天气预测。

在人们的日常生活中，智能手机中就有很多传感器，比如三轴陀螺仪、加速感应器、距离感应器、环境光感应器等。借助这些传感器，可以开发出非常丰富的移动应用，比如健康运动监测、重力感应游戏等。

物质世界是人类社会生存和发展的基础，而各种物联网技术则将人与物质世界连接起

来，使得人类能够更加客观地了解自然环境，能够高效地获取“物”的信息，物质世界中的万事万物，也将和人类一样有一个唯一的身份证，通过这个唯一的身份证，将人与物质世界融为一体。

物联网技术的广泛应用有力地推动了大数据产业的发展。从规模来讲，世界上“物”的规模要比“人口”的规模大得多。自然环境以及人类生产生活的环境中有数不清的“物”，何况地球仅仅是宇宙中的一颗小行星而已，因此，物联网中采集的数据要比人类社会中采集的数据多得多，形成的数据规模也要大得多。

6.5.5　电子商务

商品交换促进了更加专业化的社会分工，提高了社会生产的总效率，大大促进了经济社会的发展。

集市、商场、超市等是实体形式的商品交易场所，商品交换的实现方式是卖家首先从供应商处采购商品，然后将部分商品作为样品摆放到货架或者摊位上，等待买家挑选和购买，买家选择并付款后将商品带走。

电子商务的出现，改变了传统商品交易的形式。在传统商务模式下，对于实物形式的商品，从商家生产出来到最终买家手中，要消耗很多的人力物力，经过多次地点的转移。而电子商务则改变了商品交易的过程，通过整合和共享商品信息，让商品信息变得公开透明，消除了供需双方的信息不对称，使得交易更加公平，有利于消费者买到物美价廉的商品。

概括起来，电子商务与传统商务主要有以下几个不同，业界称之为电子商务的三要素。

第一要素是信息流。在传统商务中，买卖双方的信息是非常不对称的，由于空间限制，买家通常只能从有限的市场中获得商品信息，包括商家信誉、商品质量、商品价格高低、商品口碑等。电子商务借助电子商务平台将所有商品信息公开在互联网上，对于买家而言，所有商品的信息都是透明的，没有什么不对称。这样，买家就可以在全范围内对比商品和挑选商品了。

第二要素是物流。一件实物商品在传统商务中要完成一次完整交易，需要在生产厂家、批发商、零售商等多个环节的不同地点之间挪动，这些操作不仅不增值，而且还会延长商品交付时间。本质上，卖家仅仅是通过让渡商品的使用价值获得商品的价值，至于商品在

多个环节的挪动是由于供给方和需求方不能直接匹配引起的。在电子商务的第一要素提到，电子商务为供求双方提供了一个信息透明的平台，那么买家完全可以首先在电子商务平台上确定需求，然后卖家再根据买家需求直接将商品交付到用户手中，而不必经过那些既增加成本又降低效率的商品挪动环节。卖家到买家的实物交付称为物流。

第三要素是资金流。在买家提交商品购买需求后，能否快捷支付成为限制交易效率的关键环节。如果存在退换货问题，还需要退款或者调整付款金额的操作，从而降低商品交易的效率。此外，买家付款到卖家实际收款之间存在时间差，每次交易的金额小，但是交易数多，因此，交易过程形成的现金流为金融机构带来了大量的利息收入。

据商务部《电子商务报告》发布：2013 年，我国电子商务交易总额 10.5 万亿元，五年来翻了两番。其中，网络零售交易额超过 1.85 万亿元，占社会消费品零售总额的比重为 7.8%。2013 年我国已经超过美国，成为全球最大的网络零售市场，在全球网络零售市场份额中占 23.9%。2014 年，我国电子商务交易总额增速为 28.65%，移动购物市场交易规模达到 8956.85 亿元，年增长率达 234.3%，农产品电子商务交易额达 870 多亿元。

在电子商务物流方面，2013 年 5 月 28 日，知名电子商务公司阿里巴巴，联合三通一达（申通、圆通、中通、韵达），宅急送、汇通，以及相关金融机构等，宣布共同组建中国智能物流骨干网（俗称菜鸟网络），成为基于互联网思维的物流模式的新尝试。京东商城的“211 限时达”服务，即以每日 2 个 11 点钟作为时间分割点进行快速投递服务，体现了我国电子商务物流系统已经具备非常高的交付能力。

在电子商务资金流方面，以阿里巴巴的余额宝最为典型。据天弘基金发布的《余额宝一周年大数据报告》（统计区间：2013 年 5 月到 2014 年 5 月），余额宝平均每天发生 358 万笔交易，累计转入 4.96 亿次,累计消费和提现 8.1 亿。

卖家和买家的特征和行为大数据会留存在电子商务平台中，成为电子商务公司宝贵的大数据资产。

卖家特征和行为大数据包括：商家信息、商品采购记录、商品支付记录、商品销售记录、商品收款记录、客户评价信息等。卖家行为数据可以作为金融机构信用评估和提供贷款的参考依据。

买家特征数据包括年龄、性别、教育程度、兴趣爱好、购买偏好等；买家行为大数据包括平台登录/签出记录、网页浏览记录、商品搜索记录、在线咨询记录、商品购买记录、商品评价记录、商品投诉记录等。买家特征和行为数据也可以用于个人信用评估和提供贷款的参考依据。

6.6 主要内容回顾

企业发展历程犹如人生发展历程，在人生理想的指引下，经历“筑巢”、“联姻”、“孕育”、“分娩”、“培育”的修炼，终于可以大展宏图，实现“腾飞”了。

大数据是否可以帮助企业“腾飞”的检验标准是实践，比如在电信、金融、互联网行业的应用实践。

1. 大数据在电信行业的应用总结

在电信行业，随着移动互联网的飞速发展，应用商店模式的成功，产生了大量的移动用户上网记录。这些移动用户上网记录由移动通信网的网络设备记录下来，每天就有 PB 级的数据规模。

传统的关系型数据库无法满足如此大的数据规模的存取，不能解决因移动用户上网资费产生的争议问题。

基于列的分布式存储系统可以实现 IT 基础设施资源的横向线性扩展，可以满足移动用户上网记录大数据的存取要求。

内容交付网络（CDN）通过在通信网络边缘设置承载应用内容的 CDN 节点，实现了移动用户对应用的就近访问，提高了移动用户的应用访问速度。

从成本效益的角度出发，需要基于价值设置 CDN 节点，在保证用户价值的前提下，为应用提供商节约成本。同样，如果电信运营商能够从价值角度为应用提供商的 CDN 节点设置提供科学的参考依据，那么也能够增加 IDC 业务收入。判断在某区域是否应当设置 CDN 节点的关键为：移动用户到应用之间是否存在跨地域和跨电信运营商网络问题？应用价值和用户价值是否超过某个预设阈值？

基于移动用户上网记录大数据可以解决以上问题。首先，基于应用的主机 IP 和访问应用的移动终端 IP 可以推算应用部署归属地和应用访问源归属地，如果两者不一致，则存在跨网、跨区域问题；其次，应用访问流量可以作为判断应用价值高低的依据；最后，区域用户 ARPU 作为判断用户价值高低的依据。根据以上三种因素的计算结果，可以形成 CDN 节点设置的依据。

除了移动用户上网记录大数据，用户通话记录大数据也是电信运营商特有的大数据资

产。基于用户通话记录大数据，可以构建基于通话行为的社交网络。

基于通话行为的社交网络分析结果可以开发出很多创新型应用。比如，商家可以基于社交网络发现通信用户的人际圈，并实施针对性营销；公共安全管理机构可以基于社交网络辅助破案，等等。

2. 大数据在金融行业的应用总结

客户信用风险是金融企业业务管理的关键环节。

金融企业的客户类型不同，信用评估的方式和关注点也是不同的。

大中型企业贷款具有“规模大、额度大、风险大”的特点，财务能力和生产经营能力决定偿贷能力和风险敞口，例如资产负债比、现金净流量、销售净利率、市场和产品竞争力等。

小微企业贷款具有“期限短、额度小、随借随还”的特点，因此靠金融机构人工完成贷款审批是不现实的，需要采用基于大数据的客户信用评估系统辅助完成。电子商务平台上关于小微企业的 B2B 采购信息和 B2C 交易信息，可以反映小微企业的现金管理能力和生产经营能力，成为小微企业信用评估的重要输入。阿里金融面向小微企业，利用电子商务平台大数据，运用“大数定律”，在贷前评估、贷中审查、贷后监测、违规惩罚等环节实施量化和自动化处理，实现了高效率的信用评估。

FICO 评分系统是个人信用评估领域的典范。FICO 评分系统主要关注 5 个方面，即客户信用偿还历史、信用账户数、使用信用年限、正在使用的信用类型以及新开立的信用账户。国内首家大数据信用评估公司 Wecash 充分利用互联网社交大数据，实现了面向个人用户信用评估的创新。

金融诈骗对金融行业造成了严重危害，破坏了社会诚信基础。在证券行业，人们把通过未公开的信息悄悄建仓，低价买入、高价卖出，获得高额利润的行为称为“老鼠仓”。据悉，中国证监会基于大数据分析平台，对证券交易所交易数据进行实时监测，对异常情况进行报警，及时发现“老鼠仓”和“捕鼠”，成为大数据在金融领域的创新型应用。

3. 大数据在互联网行业的应用总结

近年来，互联网为广大人民提供了新闻、搜索、即时消息、电子商务等不计其数的免费（或称廉价）服务，在国外出现了雅虎、谷歌、脸书、推特、亚马逊等国际知名互联网企业，在国内出现了新浪、搜狐、百度、腾讯、阿里巴巴等知名互联网企业。互联网大大

改变了个人的生活方式。

人们在享受互联网提供的便捷、廉价服务的同时，互联网服务平台也悄悄地记录下个人的特征和行为信息。以社交网络系统为例，通过大数据分析，可以发现个人的社会关系网络及其强弱程度。对于个人，可以发现多年不联系的同学、老乡、战友、同事或者具有共同兴趣爱好的朋友等；对于企业，可以通过社交关系分析发现所需的专业人才；对于公共安全管理部门，可以通过社交关系分析预测重点关注人员的犯罪倾向，并提前采取措施。

电子商务消除了商品交易过程中的信息不对称现象，有效地匹配了市场需求和供给，提高了商品购买的便捷性，因此又被称作无摩擦经济、眼球经济、新经济。

电子商务平台记录了商品交易信息，包括商家的采购、库存、销售、服务等记录以及买家浏览、搜索、选择、购买、咨询、投诉、建议等记录，形成了电子商务大数据。

电子商务大数据具有多种用途。比如，可以用于金融机构对商家和个人的信用评估，可以分析购买行为特征进行产品推荐，可以根据用户评论改进产品，可以根据商品搜索关键字确定商品采购内容，可以根据客户浏览行为优化商品购买流程等。

第 7 章
框架体系：以不变应万变

企业既要适应外部不断变化的环境，又要协调好内部的各种资源，以最节省成本的方式为用户提供最好的产品和服务，这样才能够在市场竞争者处于领先地位，在激烈的市场竞争中得以生存和发展。

企业需要根据外部市场要求，制定长远的发展战略，不断调整发展方向和重点，而发展战略需要贯彻在企业生产经营的每一个环节，需要将业务需求落地到 IT 系统之中，可见企业管理是一个复杂的系统化工程。

为了解决复杂的企业管理问题，业界提出了企业架构的思路与方法。一方面，企业架构有效地承接企业的发展战略，另一方面，企业架构又能够将企业发展战略的目标要求反映到企业日常的运营活动当中。可见，企业架构位于企业发展战略和企业日常运营之间，起到桥梁和纽带的作用。

在多种企业架构方法论之中，以 Zachman 企业架构框架最为典型，此外还包括开放组架构框架（The Open Group Architecture Framework，TOGAF）、集成式架构框架（Integrated Architecture Framework，IAF）、美国首席信息官协会（National Association of State Chief Information Officers，NASCIO）等。

大数据运营体系需要在科学的、经过实践检验的方法论的指导下构建，为了保证本书内容具有清晰的逻辑和严谨的体系架构，本书参考了两个均具有 20 余年发展历史的框架体系。一个是电信管理论坛（TMF）推出的 Frameworx 框架体系，另一个是英国政府商务部（OGC）推出的 ITIL 框架体系。

Frameworx 框架体系为电信价值链的各参与方（电信运营商、设备供应商、服务提供商、软件开发商、系统集成商等）提供了一个公共参考框架。Frameworx 框架体系包括业务框架、信息与数据框架、应用框架以及系统集成框架，旨在为电信行业构建一个和谐的生态环境。

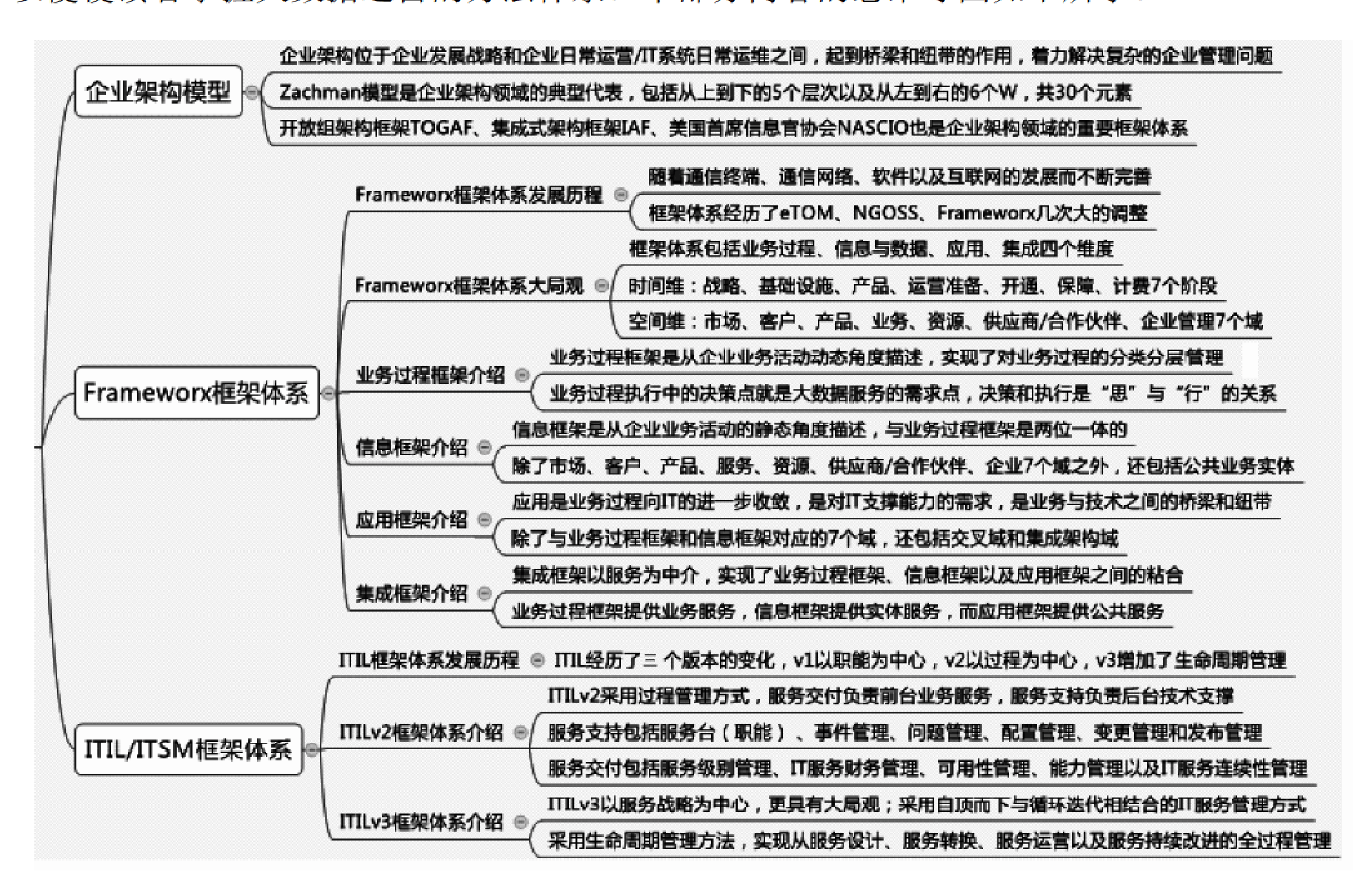
由于 Frameworx 框架体系立足于电信运营，因此更侧重于对业务框架和应用框架的刻

画，对于电信运营支撑系统的落地实施，则通过集成框架和技术中立框架予以屏蔽。因此，本书主要参考 Frameworx 框架体系中的业务部分以及“业务->应用（能力）->技术”的关联映射思想。此外，笔者对多年的信息系统规划设计与工程项目实施经验进行提炼总结，从 10 个视角（业务过程、信息、应用、集成、功能、数据、技术、部署、安全、治理）定义了企业在不同层面、不同阶段的框架设计方法及方案。

大数据服务属于企业架构中支持决策的部分，大数据服务与负责执行的操作型应用相互配合，共同完成企业在高层战略、中层管理、基层执行三个层次的业务活动。

企业架构蓝图对企业提出了 10 个维度的目标要求，但还是无法保证大数据服务的落地实施，还需要按照软件工程的思想，将大数据服务从创意转换为实现。ITIL 框架体系以服务战略为中心，参考软件工程的瀑布模型和循环迭代的设计思维，可以作为大数据服务落地实施的方法论指导。ITIL 将 IT 服务划分为设计、转换、运维、持续改进 4 个闭环的阶段，以过程管理为导向，实现 IT 服务的全生命周期管理。大数据服务是一种 IT 服务，同样可以沿着设计、转换、运维、持续优化的 IT 服务管理思路，通过持续的运营，为企业提供更加及时、有效的决策支持服务。

下面分别介绍 Frameworx 框架体系和 ITIL 框架体系的发展历程、实现方法与思路，以便使读者掌握大数据运营的方法体系。本部分内容的思维导图如下所示。



7.1 企业架构：战略与运营之桥

从不同层次、不同视角刻画企业，形成既能够承接企业发展战略，又能够指导企业日常运营的企业架构框架。

企业需要根据外部环境的变化不断调整发展战略，而企业发展战略需要贯彻到运营活动和 IT 系统之中才行，而企业架构能够在承担这一中间角色中发挥关键的作用。

20 世纪 80 年代，IBM 公司员工 Zachman 提出了“信息系统架构框架”的概念，此外还有欧共体总体框架的 TOGAF、联邦总体架构框架的 FEAF 等。

Zachman 虽然首次提出信息系统架构框架的概念，但是还是从企业建模的角度出发，多局限于信息系统架构，随着时代的发展，业界才明确提出了企业架构的概念。

Gartner 对企业架构的定义为：企业架构是能够对破坏性外力做出主动、全面、及时响应的原则，它依靠识别和分析变革执行的效果来完成。企业架构价值的交付是通过业务和 IT 部门共同签字认可后调整策略和项目实现业务目标的。企业架构用于引导决策朝着未来的目标架构演进。

维基百科对企业架构的定义为：一个定义明确的实践，用于引导企业的分析、设计、规划以及实现，每时每刻都采用一种全局思维以保证战略开发与实施的成功完成。企业架构采用架构原则和实践来指导组织，贯穿了业务、信息、流程以及技术的必要变化到战略的执行所有。这些实践是站在企业的不同视角来识别、激励并实现这些变化的。

企业架构在企业发展战略和企业日常运营/运维中的定位如图 7-1-1 所示。

从图 7-1-1 可以看出，企业发展战略分为业务发展战略和 IT 发展战略两个部分，两者是相辅相成的关系，企业发展战略决定 IT 发展战略的目标、方向和重点，而 IT 发展战略反过来也制约着业务发展战略的制定。

在企业发展战略的指导下，需要完成企业的架构设计，包括治理架构、业务架构、IT 架构以及转换架构。治理架构负责确定企业架构的治理规则，保证企业架构能够顺利地实施。业务架构确定企业业务发展的内容，包括业务过程、业务规则、信息模型等。IT 架构

确定 IT 系统的能力蓝图、功能架构、数据架构、技术架构、集成架构、部署架构等，确保 IT 发展战略的落地实施。转换架构负责将企业架构转换为企业日常运营状态或者 IT 系统运维状态，包括业务培训、IT 系统使用培训、知识管理等。

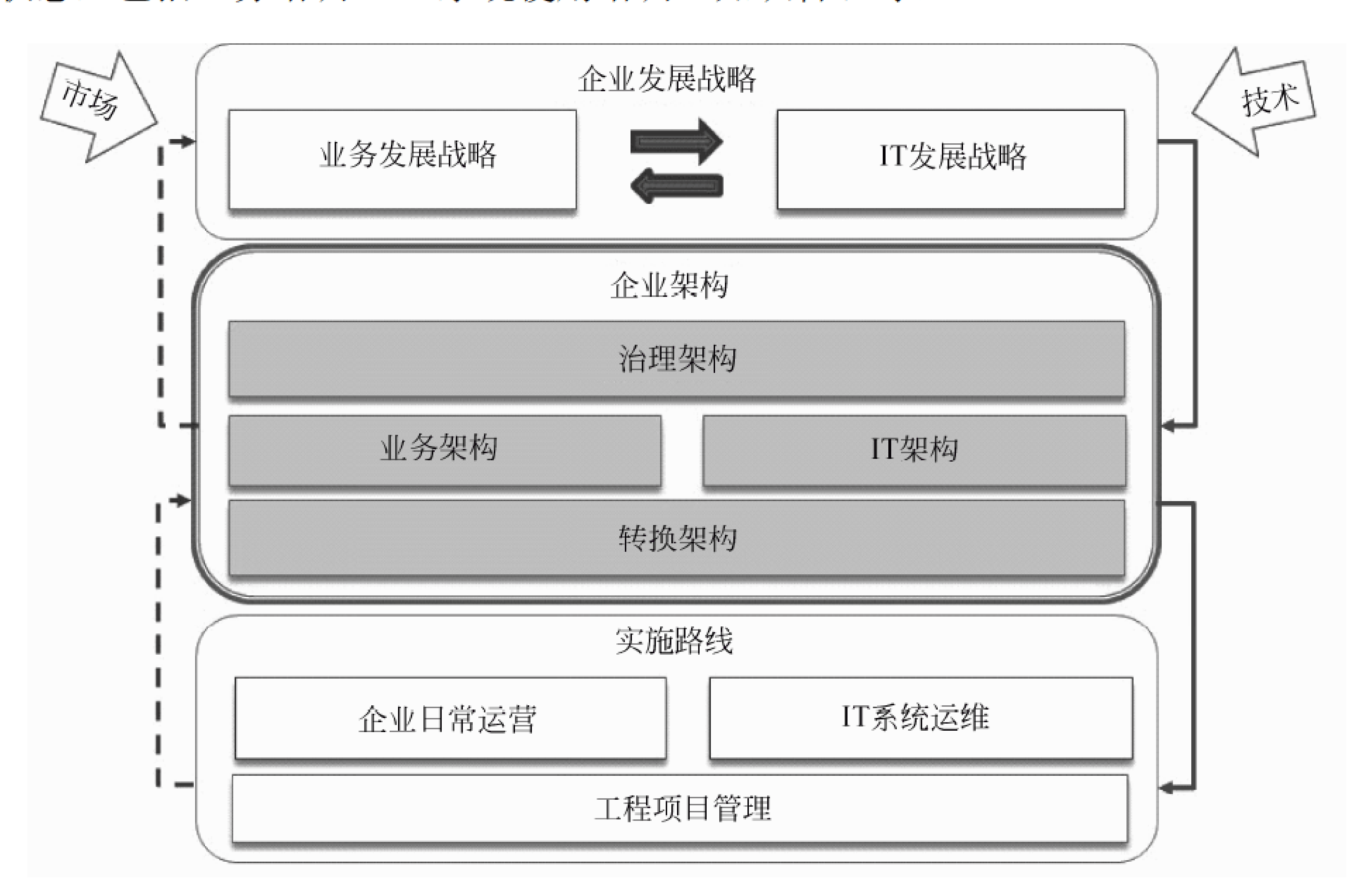


图 7-1-1　企业架构的定位

为了制定能够有效衔接企业发展战略和企业日常运营的企业架构框架体系，业界提出了许多企业架构模型，其中以 Zachman 企业架构框架最为典型。Zachman 企业架构框架模型采用 5 行 6 列共 30 个元素的矩阵式设计方法，如图 7-1-2 所示。

从图 7-1-2 可以看出，Zachman 企业架构框架采用 5 行 6 列，共 30 个元素的矩阵式方式进行设计。

从横向行维度看，其采用自上而下逐步落地的分层方式，Zachman 企业架构框架分为 5 行，即业务范围（Scope）、业务模型（Business Model）、系统模型（System Model）、技术模型（Technical Model）以及详细描述（Detailed Description）。业务范围定义了系统在功能、成本等方面的整体性要求，对应参与方是规划设计人员；业务模型描述业务流程、业务实体以及实体之间的关系，对应参与方为企业业务人员；系统模型描述系统功能和数据模型，对应的参与方为系统设计人员；技术模型定义系统开发的技术方案、平台、工具等，对应参与方为技术设计人员；详细描述定义系统的功能模块、数据库、开发接口等，

保证能够分配给开发者任务，对应的参与方为开发人员。

Zachman 框架	数据（What）	功能（How）	网络（Where）	人员（Who）	时间（When）	动机（Why）	参与方
业务范围							规划人员
业务模型							业务人员
系统模型							系统设计人员
技术模型							技术设计人员
详细描述							开发人员
具体实现示例	数据、信息	功能、结构	网络、部署	组织、人员	计划安排	策略、规则	

图 7-1-2　Zachman 企业架构框架模型

从纵向列维度看，Zachman 框架认为一个系统的建设需要 6 个方面的信息，称为 6W，它们分别是数据（What）、功能（How）、网络（Where）、人员（Who）、时间（When）、动机（Why）。

Zachman 企业架构框架模型从不同层次、不同视角、不同关注点，多方位、全面系统地定义了企业架构的内容，成为企业架构框架方面的权威指南。除了 Zachman 企业架构框架，业界还有一些典型的企业架构框架，比如开放组架构框架（The Open Group Architecture Framework，TOGAF）、集成式架构框架（Integrated Architecture Framework，IAF）、美国首席信息官协会（National Association of State Chief Information Officers，NASCIO）等。

开放组架构框架（TOGAF）包括业务架构、数据架构、应用架构、技术架构四个部分。集成式架构框架（IAF）分为业务、信息、IT 系统、基础技术 4 个部分，此外还包括管理和安全两个通用部分，IAF 又将各个组成部分分为四个层次，即环境层次（Why）、概念层次（What）、逻辑层次（How）和物理层次（With What）。美国首席信息官协会（NASCIO）

将企业业务架构分为信息（What）、功能（How）、地点（Where）、人员（Who）、业务周期（When）和业务动力（Why）几个部分。

可见，无论是哪个企业架构框架，与 Zachman 企业架构框架都具有非常相似之处，其目标都是通过从不同层次、不同视角刻画企业的方式，形成既能够承接企业发展战略，又能够指导企业日常运营的企业架构框架。

7.2 Frameworx 框架体系：电信行业的灯塔

业务过程框架、信息框架、应用框架、系统集成框架从四个不同视角定义业务、能力以及业务服务需求，为四位一体的框架体系架构。

自从伟大的贝尔先生发明电话以来，通信技术从模拟到数字，从有线到无线，从电路交换到分组交换，不断突破时间和空间的限制，取得了一个又一个成就，其发展历程如图 7-2-1 所示。



图 7-2-1　通信终端、核心网以及无线网发展历程

从图 7-2-1 可以看出，无论是通信终端还是通信网络，均发生了很大的变化：从模拟信号到数字信号，从电路交换到分组交换，从控制与承载一体到控制与承载分离，从单一功能设备到智能终端，从无线网络 Byte 级别的传输速率到百 MB 级别的传输速率，相当于最初传输速率 10^8 的级别。

信息通信技术的飞速发展促进了整个产业链的发展，与此同时，对于信息通信业务的管理日益复杂，因此，如何对信息通信产业链进行有效管控成为一个非常关键的问题。

7.2.1 Frameworx 的发展历程

随着通信技术在社会生活中的广泛应用，作为提供通信业务的承载网络也变得越来越复杂，因此电话的发源地美国牵头成立了网络管理论坛（Network Management Forum，NMF），NMF 的目标是发动通信行业内的各个参与方，制定一个各参与方共同遵循的网络管理参考框架，以推动整个通信行业的发展。

随着整个通信行业的不断发展，行业内部的专业化分工越来越细，出现了软件开发商、服务提供商、系统集成商等新的参与方，同时 NMF 的管理对象也开始从网络管理拓展到业务管理，为了适应新的发展要求，NMF 更名为电信管理论坛（Telecom Management Forum，TMF）。电信管理论坛/网络管理论坛的发展历程如图 7-2-2 所示。

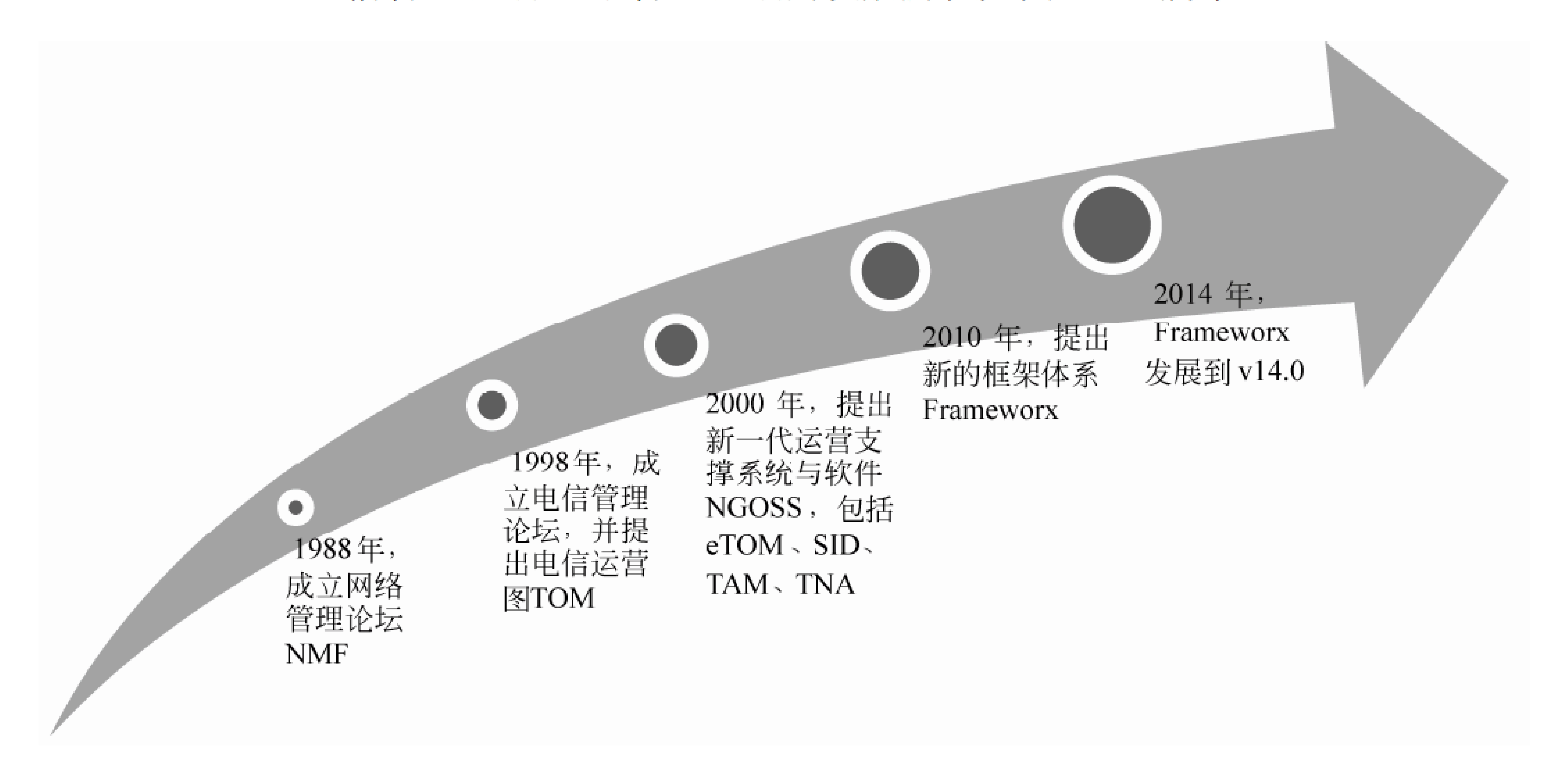


图 7-2-2　电信管理论坛/网络管理论坛的发展历程

经历了二十余年的发展，电信管理论坛形成了一套面向电信行业的、结构严谨的方法体系。

随着电信产业链的日益成熟，产业链内部的专业化分工越来越细，出现了更多的参与方，比如内容提供商、系统集成商、软件开发商、应用开发商等，迫切需要制定一个各参与方能够共同遵循的参考框架，包括业务框架、应用框架、集成框架等。各个参与方的公共参考框架需求如图 7-2-3 所示。

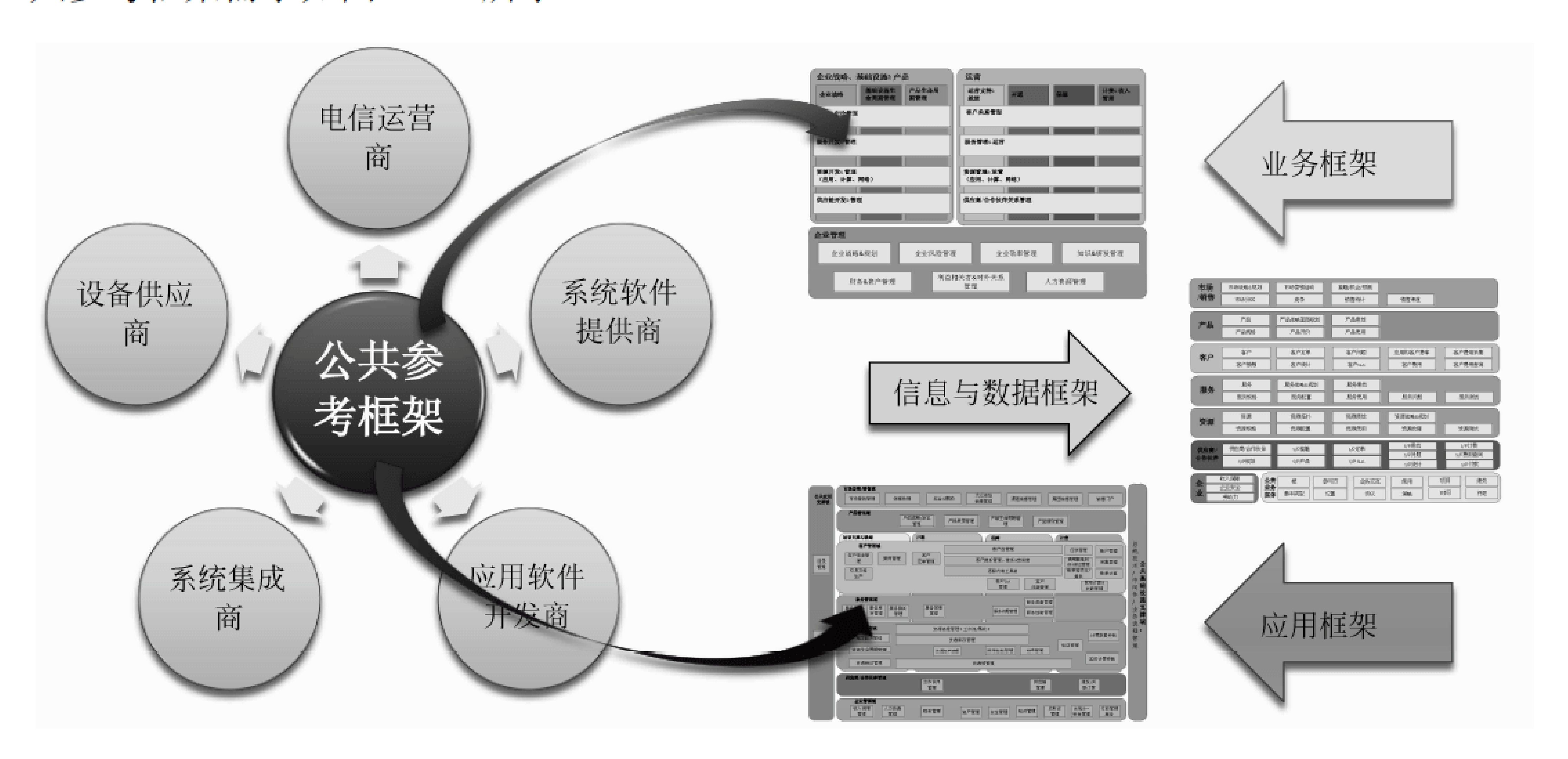


图 7-2-3　各参与方需要一个公共的参考框架

从图 7-2-3 可以看出，如果各个参与方有一个公共参考框架，那么就相当于合唱团有了一个指挥家，供给方只需按照“约定”生产，需求方只需按照“约定”采购即可。

TMF 正式成立后，首先提出了电信运营图（Telecom Operations Map，TOM），这是 TMF 从网络管理向业务管理拓展的第一个里程碑。

TOM 解决了客户视角的业务管理问题，主要是电信行业面向市场竞争的结果，是电信运营企业从“坐商”到“行商”思维转变的结果。

尽管 TOM 实现了这次伟大的转变，但是 TOM 仅限于运营管理，不能全面地观察电信企业从战略、建设、运营以及管理的全过程。为此，TMF 在 TOM 的基础上进行了丰富和完善，增加了战略、基础设施、产品和企业管理几个域，组成了一个面向电信企业完整的框架体系，称为增强的电信运营图（enhanced Telecom Operations Map，eTOM）。

eTOM 其实是 TMF 提出的新一代运营支撑系统与软件（New Generation Operations

Systems and Software，NGOSS）的一部分，除了 eTOM，NGOSS 还包括共享信息与数据（Shared Information and Data，SID）、电信应用图（Telecom Application Map，TAM）、技术中立架构（Technology Neutral Architecture，TNA），NGOSS 的目标实现电信运营支撑软件的“即插即用”。

在 NGOSS 体系中，eTOM 从企业活动出发来描述企业业务过程，SID 则从业务活动中形成的信息和数据出发描述企业数据模型，TAM 则从应用（能力）的角度出发来描述企业对于电信运营支撑系统的能力要求，以便电信产业价值链中各参与方能够有一个共同的能力参考框架，消除不同产品与服务提供商的产品重叠问题，实现更节省、更快速的系统集成。

随着互联网的飞速发展，社会分工更加专业化，全球化的资源配置与协同使得企业从简单的价值链模式进化为价值网络模式。价值网络时代对于企业运营提出了新的要求，电信运营商为了适应这一新的变化，需要调整现有架构，采用面向服务的架构（SOA）的架构模式，满足互联网时代快速协同的要求。

TMF 适应这一发展趋势，经过多次论证讨论，在 2010 年提出了全新的 Frameworx 框架体系。

7.2.2　Frameworx 框架体系大局观

Frameworx 框架体系以商业需求为输入，将运营支撑框架分为业务、应用、技术三个部分，各部分各有侧重又相互联系，形成了一体化的、贯通业务与技术的完整框架体系。

在业务层面，从业务过程和信息数据两个视角分别刻画；应用层面，将业务需求向 IT 实现进一步收敛，形成了面向多个域的支撑能力；技术层面，将原来的 TNA 调整为 SIF，通过业务服务（Business Service，BS）将业务过程、信息数据、应用三个方面有机地结合起来。业务服务也称为合约（Contract）。

x 是一个变量，可以表示多个不同的值。可见，Frameworx 是 *x* 个 Framework 的集合体。Frameworx 包括业务过程框架（对应 eTOM）、信息框架（对应 SID）、应用框架（对应 TAM）以及系统集成框架（对应 TNA）。Frameworx 框架体系如图 7-2-4 所示。

Frameworx 是 TMF 从面向服务的角度出发，对于 NGOSS 进行了重新设计得到的。Frameworx 除了继承 NGOSS 的 eTOM、SID 以及 TAM 之外，主要亮点是采用业务服务

（Business Service，BS）的方式实现系统集成。

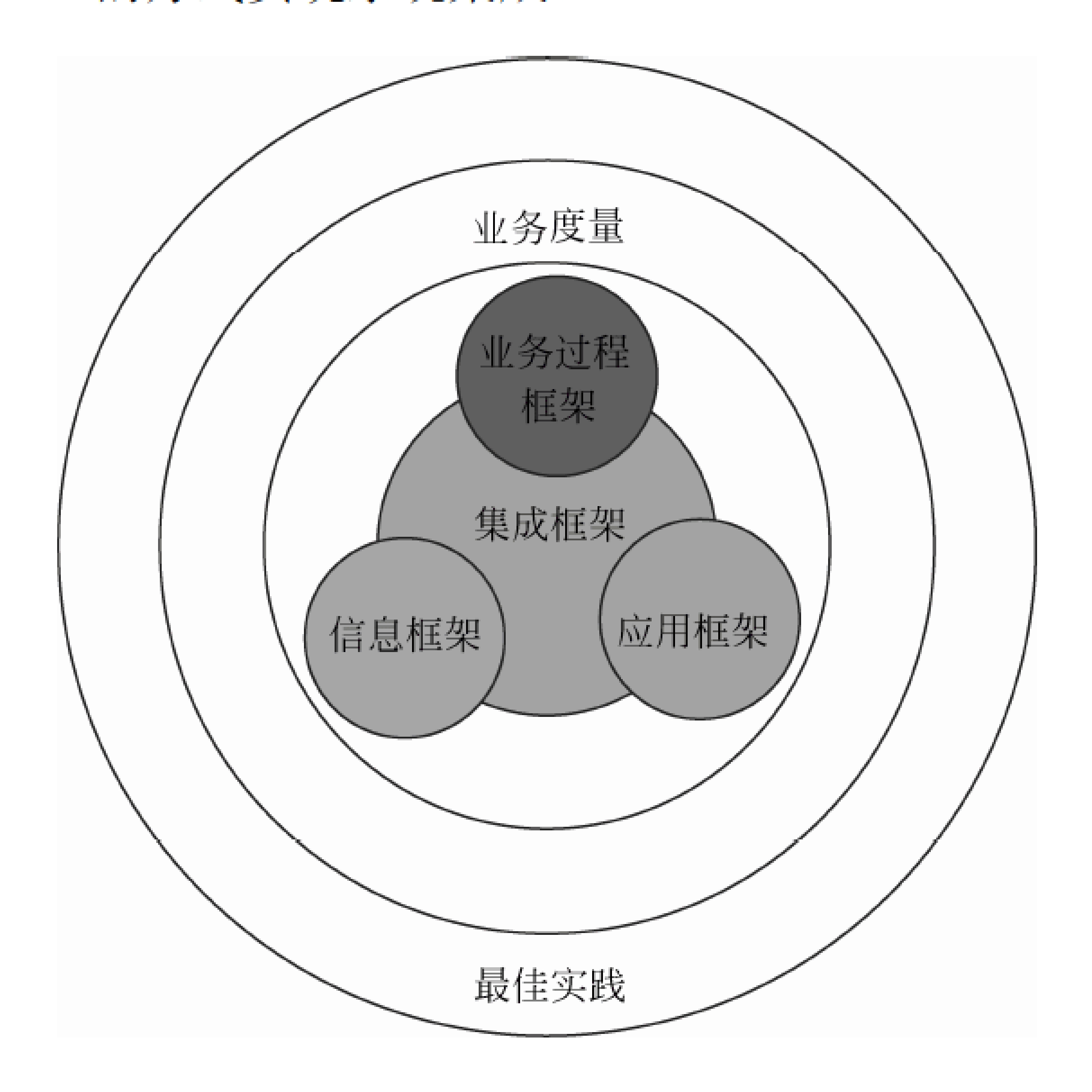


图 7-2-4　Frameworx 框架体系

业务服务其实就是采用了面向服务的思想。业务服务来源于其他几个 Framework，比如以任务为中心的服务来源于业务过程框架（即 eTOM），以实体为中心的服务来源于信息框架（即 SID），以效用（Utility）为中心的服务来源于应用框架（即 TAM）。

当根据需求定义好所有业务服务以后，可以将业务服务作为平台实现的输入，由于这些业务服务是技术中立的，因此与平台的具体实现无关，可以采用 J2EE/Java、CORBA、.NET 等任何语言和工具完成运营支撑系统的开发。

eTOM、SID、TAM、BS 的侧重点在于构建满足业务需求的框架体系，作为电信运营支撑系统实现的参考框架，还需要通过业务度量（Business Metrics）来验证其是否满足业务需求。此外，Frameworx 还根据业务和技术发展重点，给出了面向特定领域的最佳实践，通过最佳实践来解决云计算、大数据、客户体验管理安全等方面的问题。

由于 Frameworx 从业务角度来描述企业，同时又具有技术无关性，因此其框架体系具有良好的稳定性。此外，Frameworx 框架体系采用资源层与服务层分离的方式，因此可以作为其他服务型企业架构设计的通用参考框架。比如可以作为教育、医疗、交通、餐饮等服务型企业的架构设计参考。

7.2.3 业务过程框架介绍

网络管理论坛成立的初衷是制定管理通信网络的公共参考框架，随着电信业务的市场化，电信运营商不得不面对外部的市场竞争。为了满足市场需要，网络管理论坛提出了电信运营图，即 TOM（Telecom Operation Map），同时网络管理论坛（NMF）也更名为电信管理论坛（TMF）。TOM 参考框架如图 7-2-5 所示。

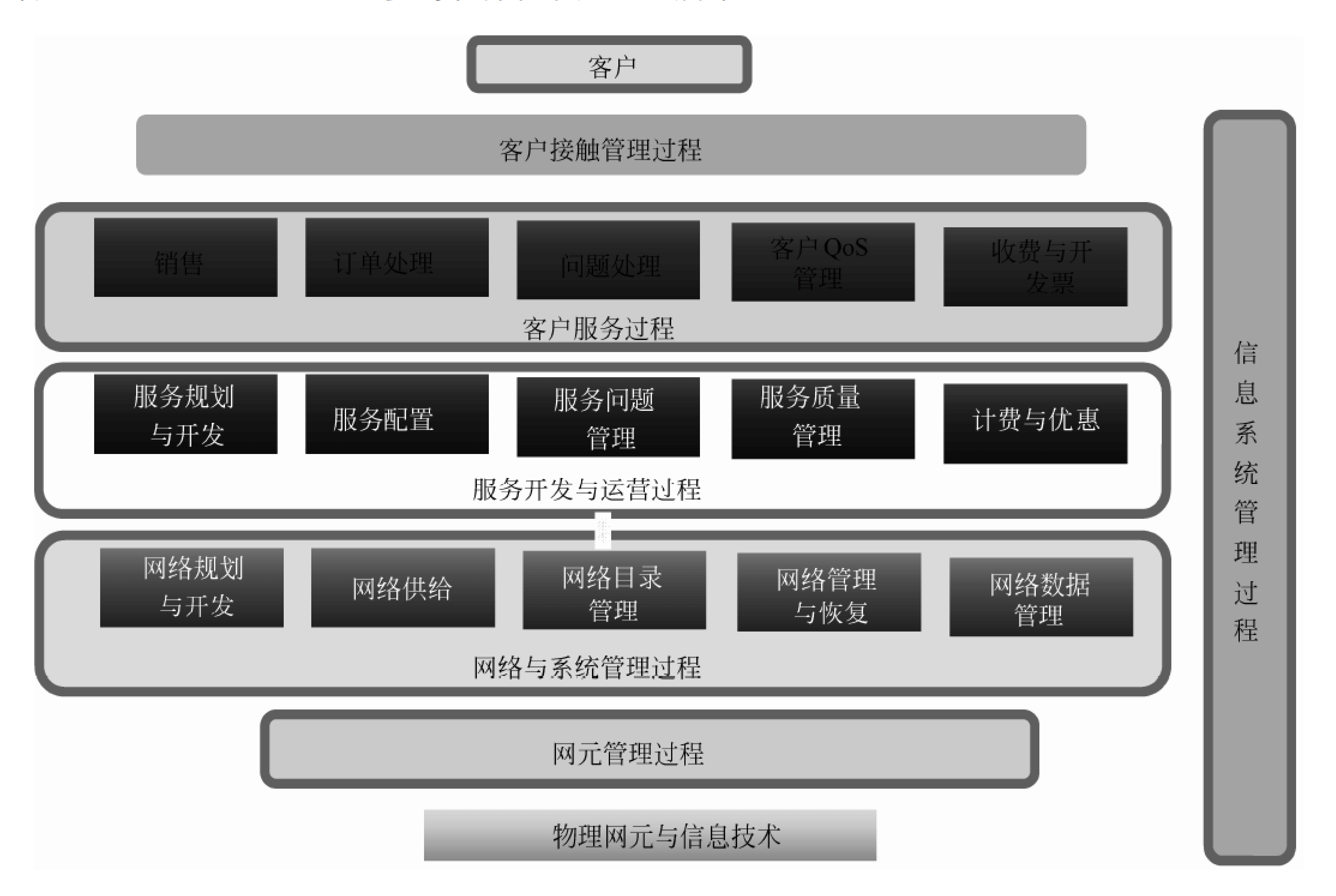


图 7-2-5　电信管理论坛的电信运营图（TOM）

电信运营图虽然将电信产品推向前台，增加了销售和客服功能，但是 TOM 还是不能立足于企业发展全局，不能全面地反映企业业务活动，为此，电信管理论坛在原来 TOM 的基础上进行了增强，制定了增强的电信运营图，即 eTOM（enhanced Telecom Operation Map）。当初的 TOM 或者 eTOM 就是现在 Frameworx 框架体系中的业务过程框架（Business Process Framework，BPF）。

eTOM 采用分层的方式从 0 级视图开始逐步细化，以便电信运营图能够更加清晰地刻

画电信运营要求。eTOM 的 0 级参考框架如图 7-2-6 所示。



图 7-2-6　eTOM（0 级参考框架）

从图 7-2-6 可以看出，eTOM 分为两类来描述电信运营过程，一类是电信运营过程中涉及的参与方，比如客户、供应商、合作伙伴、股东、雇员以及其他利益相关者，另一类是电信运营框架自身，在 0 级参考框架中分为三个相互独立的域，即战略&基础设施与产品域、企业运营域、企业管理域。

eTOM 的 0 级参考框架仅仅是一个起点，需要在此基础上对企业业务过程进一步细分，直到实际执行的业务过程块。

7.2.4　信息框架介绍

业务框架为一体两翼，一翼是业务过程框架，另外一翼是信息框架。信息框架形成的源头是企业业务过程产生的信息，而这些信息是需要概念模型来承载的。信息与业务过程

共同描述业务需求，业务过程从动态角度描述，信息从静态角度描述。

在业务需求分析阶段，通过概念模型来描述实体之间的关系。概念模型虽然可以从业务视角对需求进行刻画，但是其还需要进一步细化才行，为了对概念模型进行有效的管理，提出了信息框架。信息框架与业务过程框架相对应，同样是分为市场/销售、产品、客户、服务、资源、供应商/合作伙伴、企业，共 7 个域，此外，还有一个特殊的公共业务实体，这是其他域公用的实体对象。一级信息框架如图 7-2-7 所示。



图 7-2-7　信息框架（一级）

为了直观地看到信息框架和业务过程框架的一体两翼关系，下面对这两个框架进行对比，对比图如图 7-2-8 所示。

从图 7-2-8 可以看出，业务过程框架中的第一层（市场、产品、客户）在信息框架中被分为市场/销售、产品、客户三个独立的域，其他域，如服务域、资源域、供应商/合作

伙伴域、企业管理域，则表现为与业务过程框架一一对应的关系。此外，在信息框架中，考虑到不同域的实体对象的复用性，新增了一个公共业务实体域。

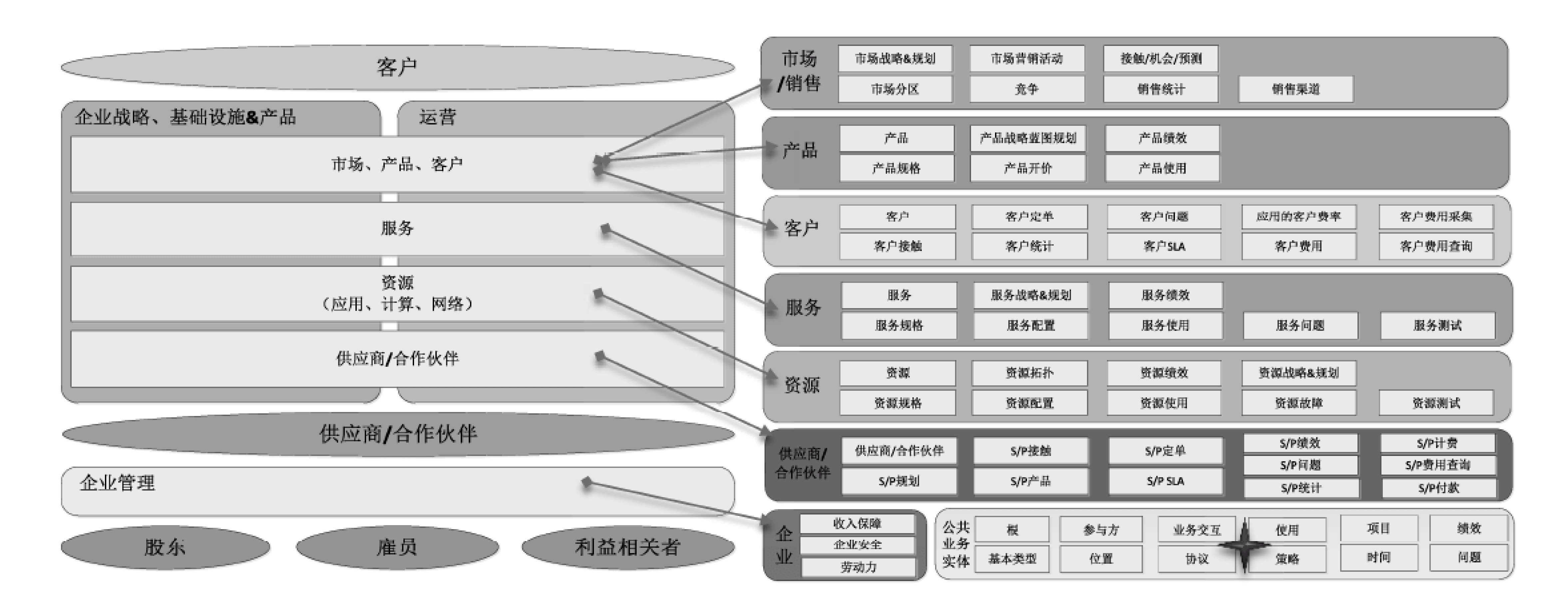


图 7-2-8 业务过程框架与信息框架对比

7.2.5 应用框架介绍

应用架构是能力的集合体，也称之为能力蓝图。应用是业务人员和技术人员之间的一座桥梁，是他们之间沟通的媒介。业务人员可以对技术人员说："你们要实现这些能力，这是我们的需求，有了这些能力，我们的业务能力就强大了！"技术人员也担心自己说的话太"技术"，业务人员听不懂，并且担心因为没有沟通好而白做了工作，于是就与业务人员确认："系统实现了这些能力就可以了吗？就满足你们的需求了吗？"业务人员回答说："是这样的！"

应用框架是业务过程框架向技术实现的进一步收敛，同时也包括了公用的应用，那是因为应用初步描述了技术特征，而技术是可以复用的。

应用框架与业务过程框架相对应，从纵向看，包括战略、基础设施生命周期管理、产品生命周期管理、运营支撑与就绪、服务开通、服务保障、服务计费，这与业务过程框架的分类是一致的。从横向看，包括市场_销售域、产品管理域、客户管理域、服务管理域、资源管理域、供应商/合作伙伴管理域和企业管理域，这些与业务过程框架也基本一致。与业务过程框架不同的是，应用框架还包括交叉域和集成架构域，前者是其他域共用的能力，

而后者则是为了实现应用之间的集成而需要的能力。一级应用框架如图 7-2-9 所示。

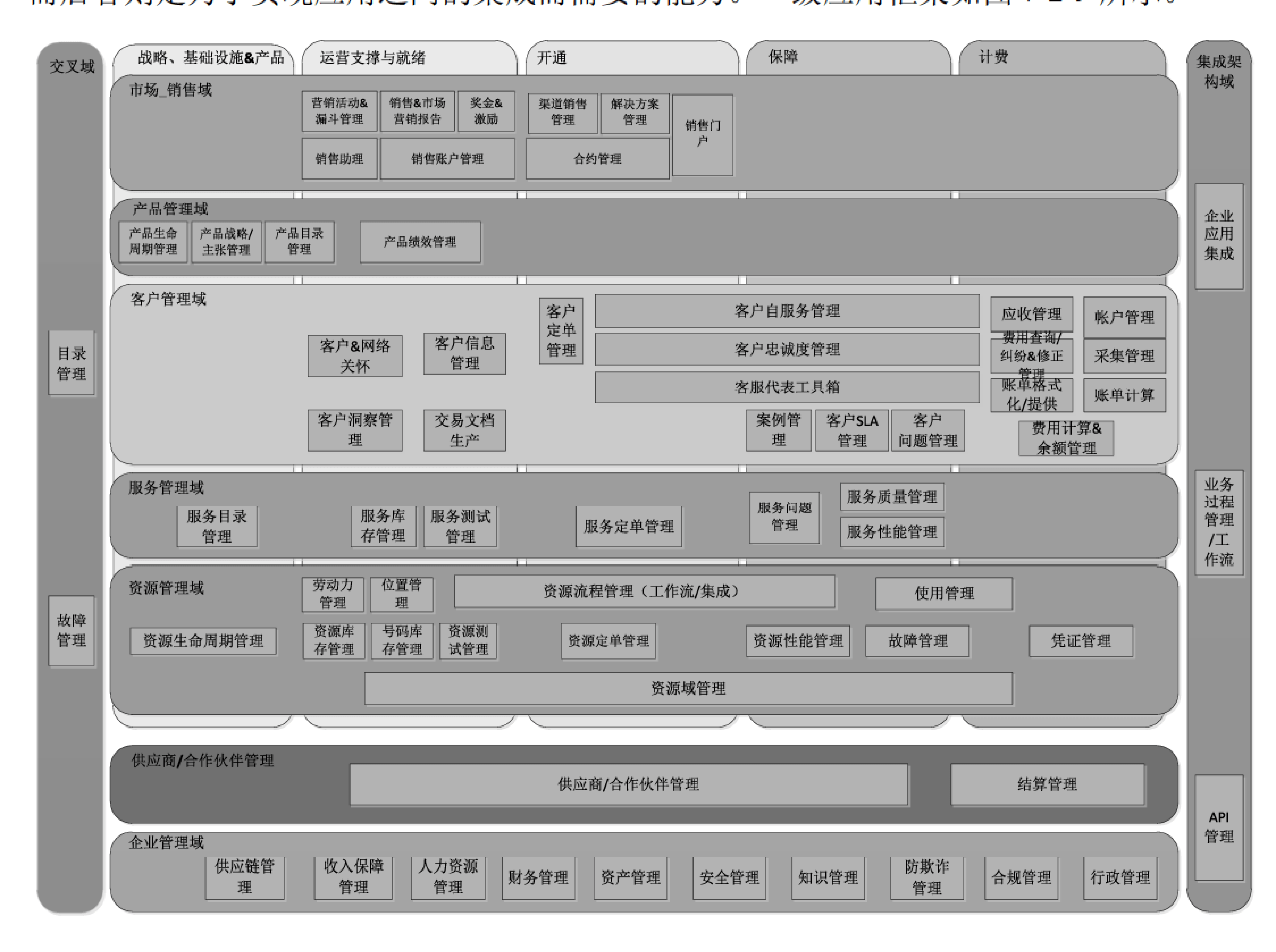


图 7-2-9　应用框架（1 级）

企业可以设计符合自身的目标能力蓝图，同时也可以分析自身能力后，形成能力现状蓝图，通过能力现状蓝图与目标能力蓝图的对比，找出企业还存在的能力差距。

业务过程框架和信息框架用于描述业务需求，但是这些业务需求最终还是需要信息系统来承载的，因此需要一个参考框架来描述业务能力。既然应用框架以业务过程框架为输入，因此它们之间势必存在着密切的联系。一级业务过程框架与一级应用框架的对比如图 7-2-10 所示。

从图 7-2-10 可以看出，业务过程框架与应用框架从战略到运营（纵向）的分类方式基本是一致的。从市场到资源的前后分层支撑来看（横向），两种存在一定的映射关系，如表 7-2-1 所示。

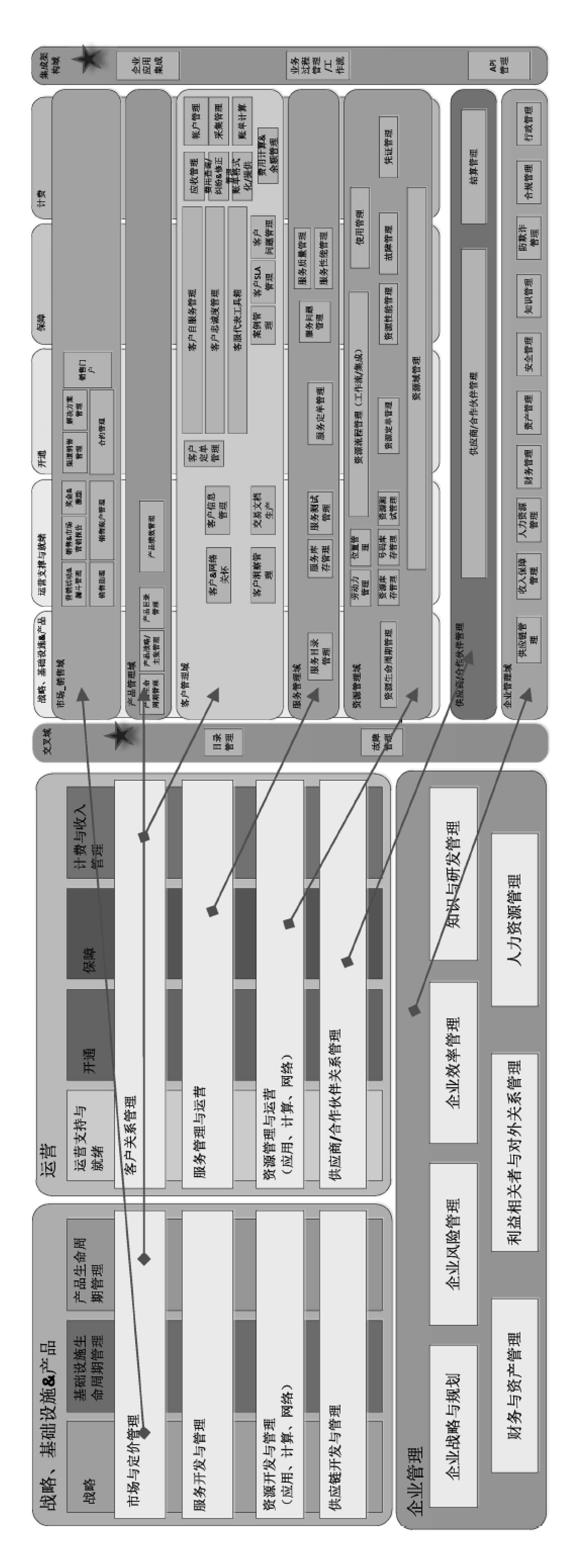


图 7-2-10 业务过程框架与应用框架对比（一级）

表 7-2-1　业务过程框架与应用框架对比（一级）

业务过程框架	应用框架对应项	解　析
市场与定价管理	市场_销售管理域-运营支撑与就绪、开通	市场销售侧重于对于销售活动的支撑，提供解决方案、合同管理等功能，目的是提升销售能力
产品生命周期管理	产品管理域-战略、基础设施、产品、运营支撑与就绪	产品首先是在战略的指导下，以基础设施为支撑而形成的，然后再完成配置、上架等工作，为正式运营做好准备
客户关系管理	客户关系管理域（除战略、基础设施、产品外）	客户关系管理完全属于企业运营阶段的事情，其完成客户的引入、服务、关怀、维系、挽留、退出等全生命周期的管理。
服务开发与管理、服务管理与运营	服务管理域（除计费外）	服务是连接市场（含客户、产品、渠道等要素）与资源的桥梁，是虚拟的，不直接产生价值，因而无须费用的计算
资源开发与管理、资源管理与运营	资源管理域（全部）	资源贯穿企业战略和运营的全过程，是企业价值创造的基础
供应链开发与管理、供应商/合作伙伴管理	供应商/合作伙伴管理域（除战略、基础设施、产品外）、企业管理域-供应链管理	供应商/合作伙伴管理类似于客户关系管理域，主要侧重对供应商/合作伙伴的准入、退出、考核、结算等方面的管理
企业管理	企业管理域（新纳入供应链管理）	企业管理域中除了人力、财务、资产、安全、风险、行政等功能外，还纳入了供应链管理，包括供应链规划、采购、运输、后勤、订单跟踪等管理功能

此外，应用框架中还引入了交叉域和集成架构域。交叉域中的应用能力可以为多个域共享，包括目录管理和故障管理。集成架构域为应用之间集成的通用型应用，包括企业应用集成（比如企业服务总线、消息总线等）、业务流程管理/工作流、API 管理。

可见，应用框架是业务过程框架的进一步收敛，同样是分域管理的，每个域对应一个能力集。运营企业可以参考应用框架进行产品和服务的采购以及应用的实施，应用软件提供商、系统集成商等参与方也可以参考应用框架，对自身产品进行定位，避免与其他供应商提供的应用产生交叉和重叠。

7.2.6　集成框架介绍

当前，社会专业化分工越来越细，作为社会生产中的每一个环节，都不可避免地与其他企业或个人进行交互，需要借助集成其他应用来实现某一个特定的业务功能。为达到既能满足业务需求又能适应技术发展变化的目的，TMF 提出了通过业务服务（即合约）实现

集成的思路，业务服务的集合就是系统集成框架的具体内容。

TMF 将业务服务定义为：在 SOA 的语境下，人工服务与自动服务的综合体，实现特定的业务功能或特性，提供业务能力访问的途径。此外，业务服务采用价值链方法，在声明自身提供的服务能力的同时声明其所依赖的服务，如图 7-2-11 所示。

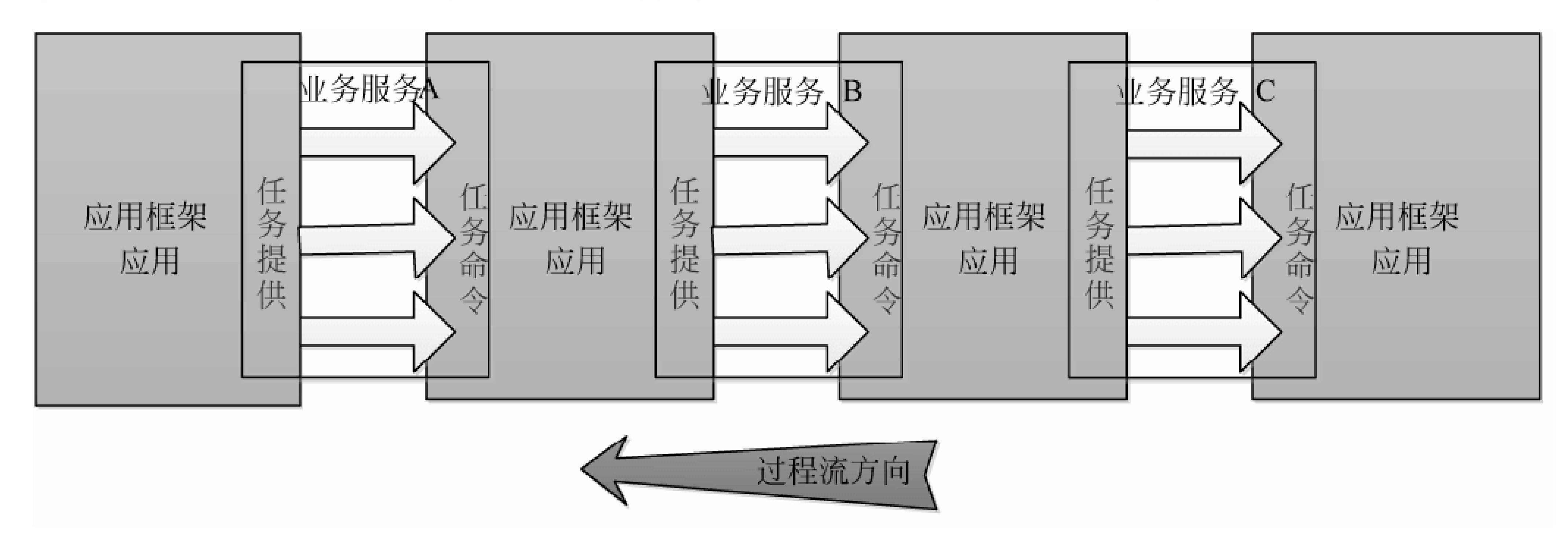


图 7-2-11　价值链思维的业务服务（合约）

业务服务在业务过程框架、信息框架、应用框架的连接关系中承担黏合剂的角色，在业务需求到技术实现的过程中所处的位置如图 7-2-12 所示。

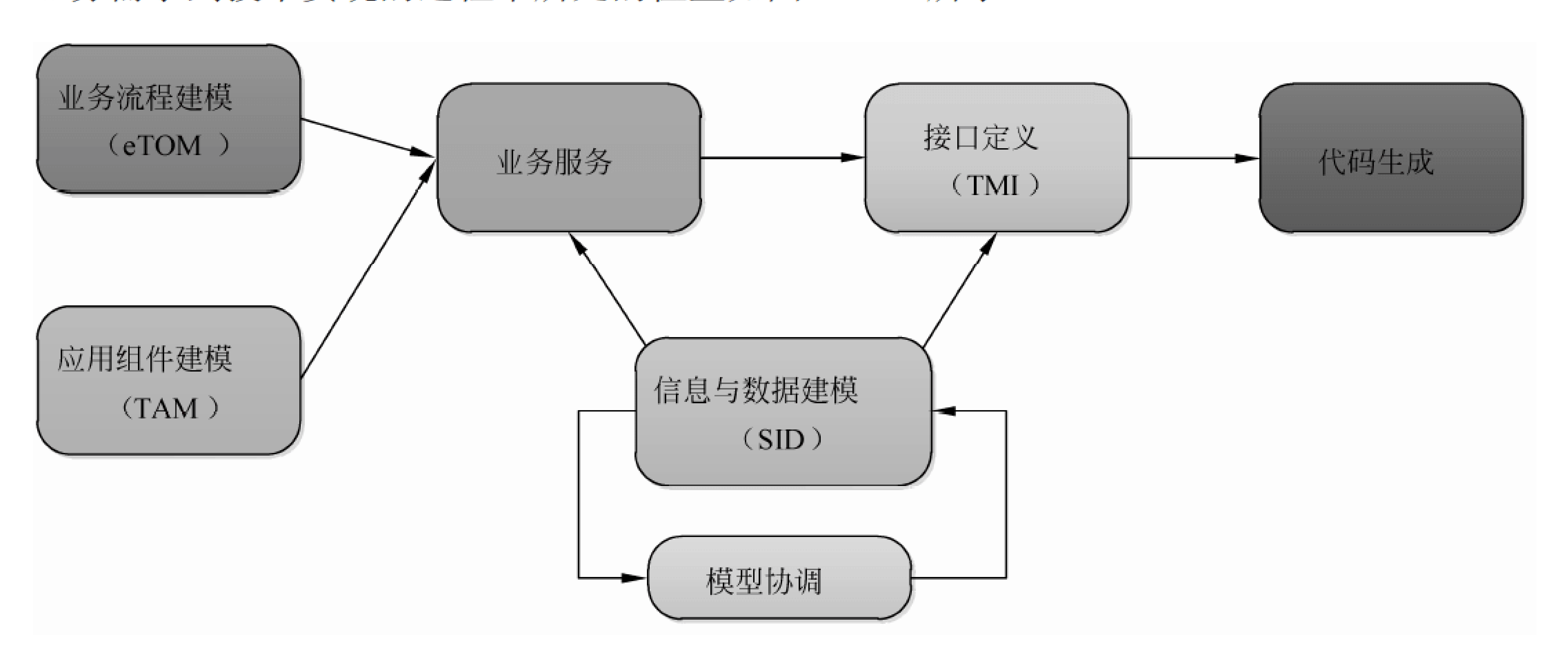


图 7-2-12　业务服务（合约）创建过程

从图 7-2-12 可以看出，业务服务以业务过程建模（提供操作信息）、应用组件建模以及经过调整后的信息与数据建模（提供属性信息）为输入，构造出各种业务服务，然后在此基础上进行接口定义，最后根据定义的接口生成代码框架。TMI 与接口定义语言（IDL）

类似，IDL 提供通用数据类型，是实现跨平台的基础。

7.3　ITIL/ITSM 框架体系：IT 行业的指南针

以服务方式管理 IT，采用全生命周期的管理方式，分为服务战略、服务设计、服务转换、服务运营、服务持续优化 5 个阶段。

在 IT 治理方面，由 OGC 发起并形成的国际规范 ITIL（IT 基础设施库）最为典型。顾名思义，ITIL 的管理对象是 IT，而 IT 又是以服务的形式提供给使用者的，因此对于 IT 的管理又称为 IT 服务管理，即 ITSM（IT Service Management）。

从管理范围角度看，ITIL/ITSM 的管理对象包括应用软件和基础设施两个层面，如图 7-3-1 所示。

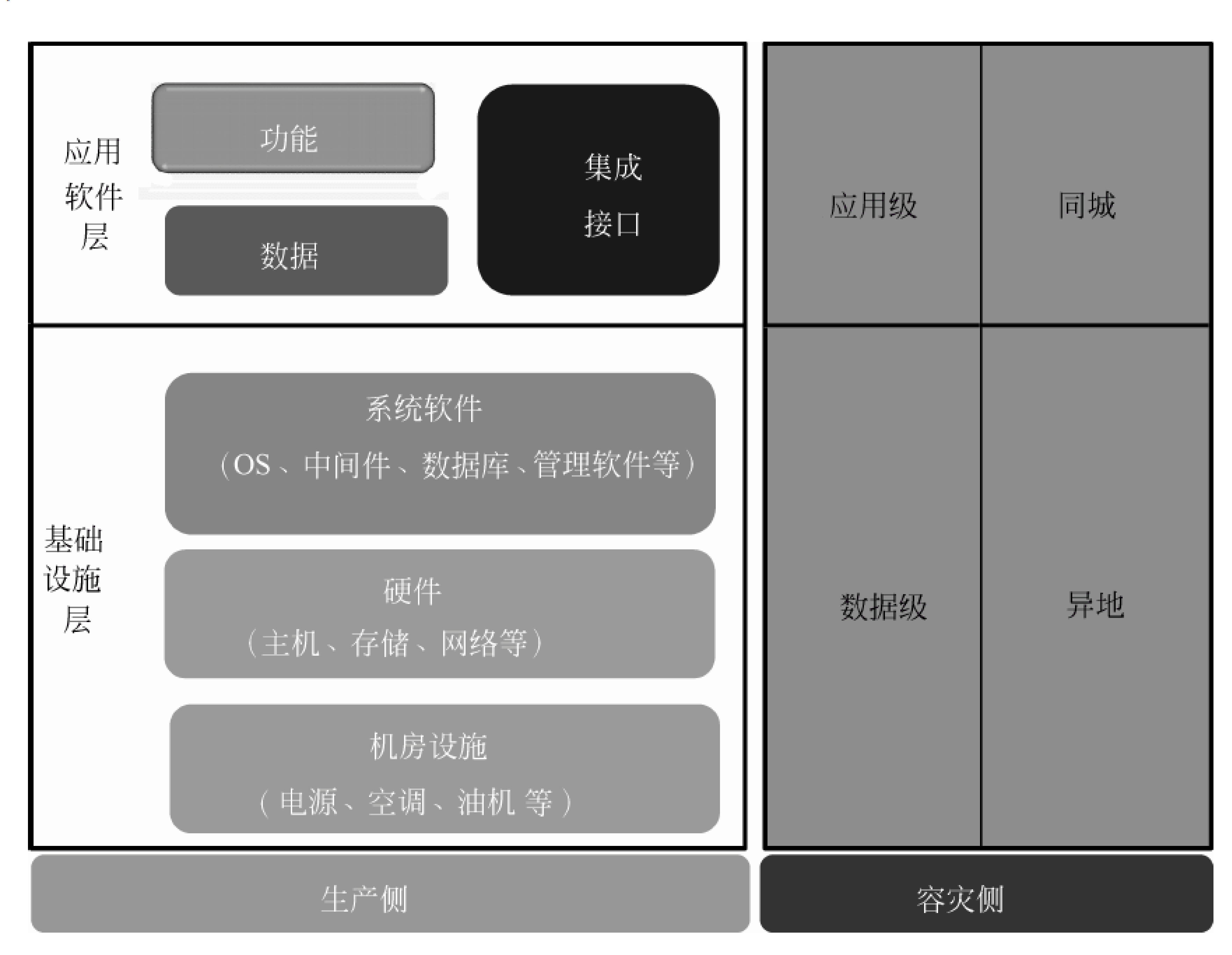


图 7-3-1　ITIL/ITSM 的管理范围

从图 7-3-1 可以看出，ITIL 负责对应用层和基础设施层的软件和硬件设备的管理，以

保障 IT 服务能够满足用户的正常使用。从 IT 系统的功能看，分为支撑组织生产的基础设施和应用，也有保障生产系统可靠性的容灾系统。

生产侧 IT 系统分为基础设施层和应用软件层，基础设施层又分为机房设施、系统硬件和系统软件三层，应用软件层包括功能、数据、集成接口三个部分。

与生产侧 IT 系统相呼应，容灾系统从保障级别角度分为数据级容灾和应用级容灾，容灾系统从保障范围角度，分为异地容灾和同城容灾。

7.3.1 ITIL/ITSM 框架体系发展历程

ITIL 产生于 20 世纪 80 年代，到目前为止，已经经历了三个版本的发展演进。ITIL 发展历程如图 7-3-2 所示。

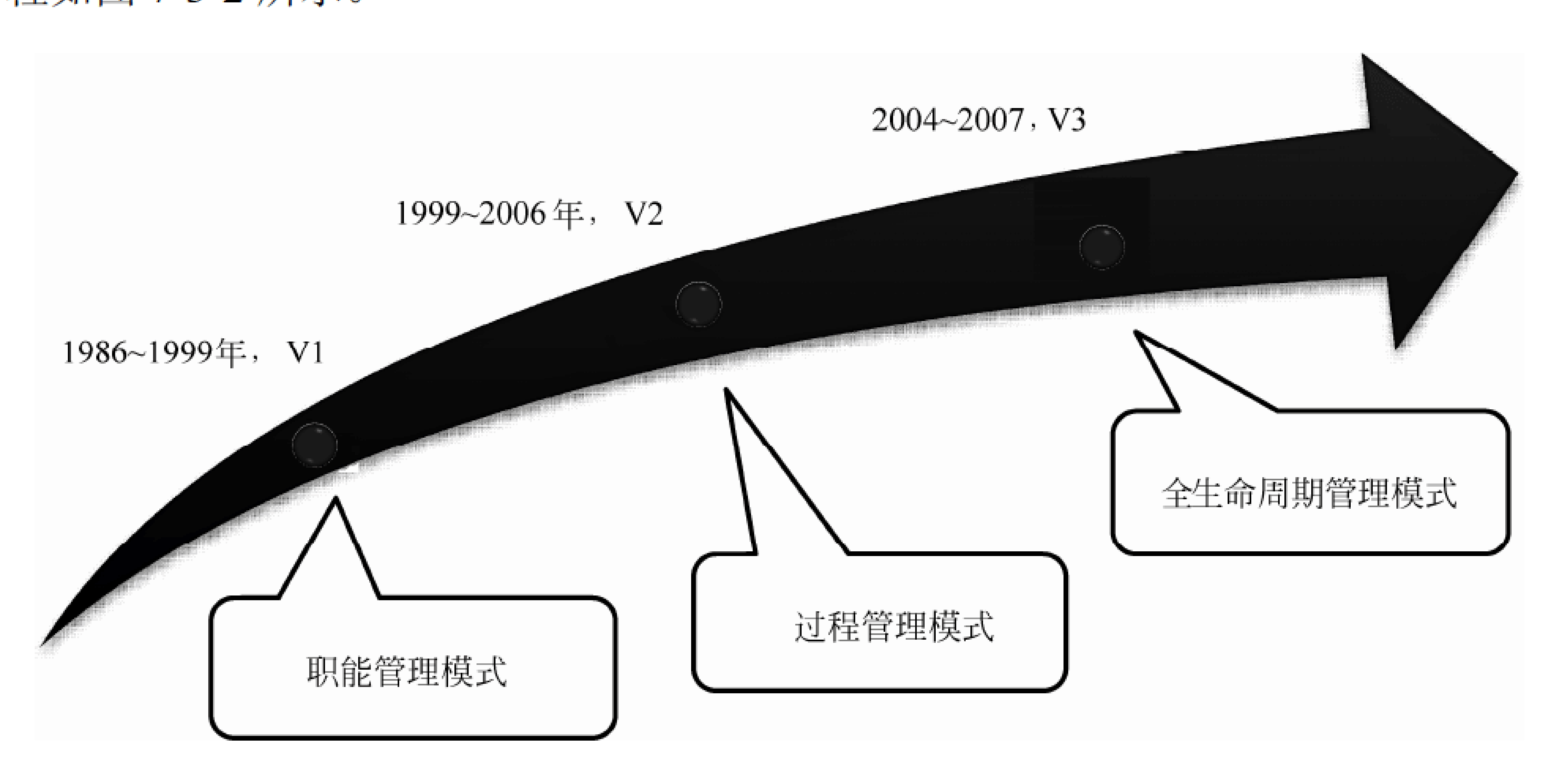


图 7-3-2 ITIL 发展过程中版本的变化

从 1986 年到 1999 年，属于 ITIL 第一个版本提出和应用的阶段，ITILv1 采用职能管理模式，主要解决组织的工作效率问题。

从 1999 年到 2006 年，属于 ITIL 第二个版本提出和应用的阶段，ITILv2 采用过程管理模式，解决了各个不同职能部门之间的协同和信息共享问题。

从 2004 年到 2007 年，属于 ITIL 第三个版本提出和应用的阶段，ITILv3 采用全生命周期管理模式，使得 IT 服务管理更具有大局观和整体性，开始按照成本效益的方式设计和评价 IT 服务。

7.3.2 ITILv2 框架体系介绍

ITIL 的第一个版本以职能为中心，目标是解决 IT 系统的支撑效率问题。IT 系统虽然提高了工作效率，但是多个 IT 系统往往存在功能重叠、数据不一致等不足，为了解决这一问题，ITIL 将治理架构从职能管理方式转变为过程管理方式，并形成了 ITIL 的第二个版本。ITILv2 采用了从业务到技术，从服务交付到服务支持的设计思路，定义了 10 个过程与 1 个职能，有效地支撑了面向特定任务的实现。ITILv2 框架体系如图 7-3-3 所示。

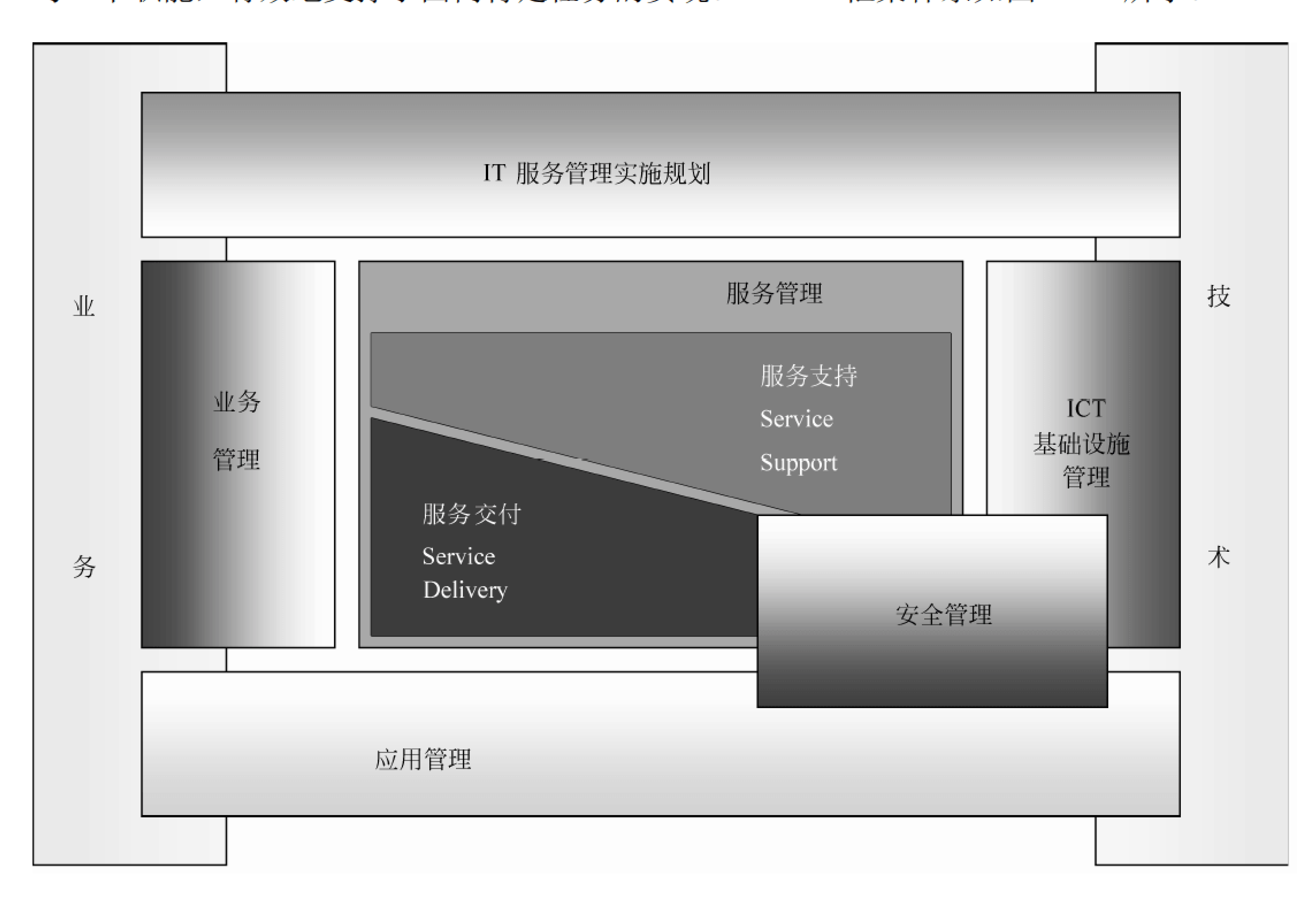


图 7-3-3 ITILv2 框架体系

ITILv2 框架体系中的两个核心模块是服务交付和服务支持。服务交付包括服务级别管理、IT 服务财务管理、可用性管理、能力管理以及 IT 服务连续性管理；服务支持包括服务台（职能）、事件管理、问题管理、配置管理、变更管理和发布管理。

7.3.3 ITILv3 框架体系介绍

ITILv2 虽然解决了 IT 系统之间的流程协同和信息共享问题，但仍旧立足于解决某个特定任务，缺乏大局观与系统性，无法全面地、全过程地计算 IT 服务的成本效益。为了解决以上不足，ITIL 提出了全生命周期管理模式，推出了 ITIL 的第三个版本 ITILv3。

ITILv3 以服务战略为指导，以服务设计、服务转换、服务运营为主线，以服务持续改进为落脚点和新的出发点，形成全新的、更加系统化的框架体系。

ITILv2 到 ITILv3 的演进路线如图 7-3-4 所示。

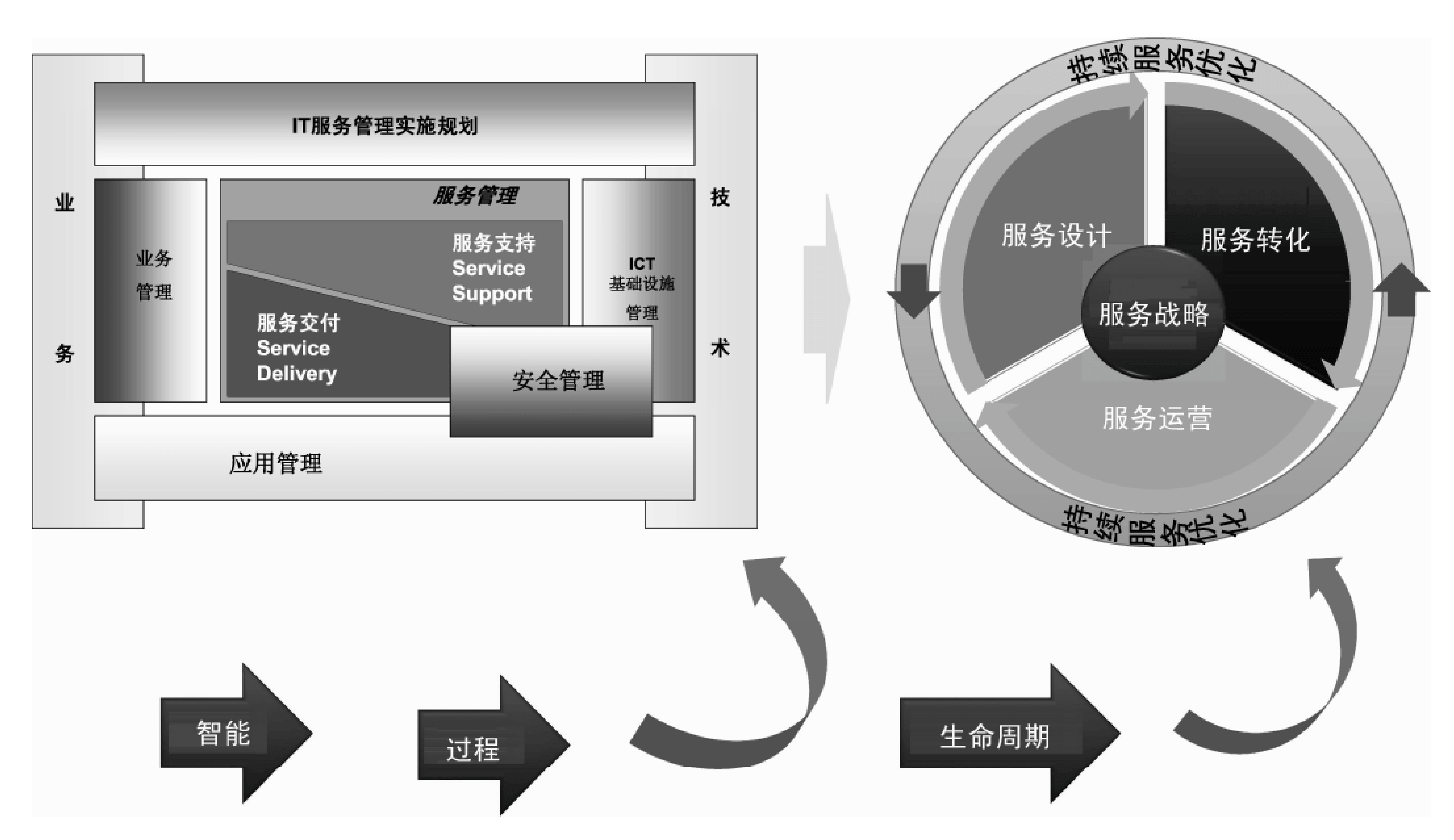


图 7-3-4 ITIL 框架体系从 v2 到 v3 的演进

ITILv3 包括服务战略、服务设计、服务转换、服务运营、服务持续优化，共 5 个阶段，形成了一个以服务战略为核心的、面向 IT 服务的、全生命周期管理的框架体系。

1. 服务战略阶段

引入服务战略思维的主要目的是实现 IT 服务更好的成本效益，使得 IT 服务像经营业务一样，能够站在客户的角度，思考 IT 服务能够为客户带来多少价值、消耗多少成本。

服务战略的制定需要遵循一定的方法和原则。服务战略制定需要从价值创造、服务资产、服务提供方类型、服务架构等角度考虑。分开来讲，价值创造是分析 IT 服务能够为客户创造什么价值、多少价值。服务资产是将资源与能力看作价值创造的基础，综合考量组织外部业务需求与内部资源能力。

在能力和资源之间取得平衡是 IT 服务管理追求的目标。能力是组织的“软”实力，包括管理、组织、流程、知识、人员几个方面，是组织对各种资源的运用；资源是组织的“硬”实力，包括财务资本、基础设施、应用、信息、人员几个方面，是组织对外提供能力的基础。能力与资源的有效匹配是 IT 服务管理的努力方向。能力与资源的关系如图 7-3-5 所示。

能力		↔	资源	
A1	管理		财务资本	A9
A2	组织		基础设施	A8
A3	流程		应用	A7
A4	知识		信息	A6
	人员	A5	人员	

图 7-3-5　ITIL 框架体系中组织能力与资源的关系

从图 7-3-5 可以看出，左侧属于能力范畴，能力体现了业务对于 IT 能力的需求，右侧属于资源范畴，资源体现了技术工具对于 IT 能力的资源支持。比如用户在使用 IT 服务的过程中，发现 IT 服务不可用，那么该用户首先是报告该问题，ITIL 记录问题并将其转入相应的处理流程，这体现了业务对于 IT 能力的需求。

仅仅具有 IT 服务的管理流程还是不够的，流程的实现还必须通过组织的资源落地，比如用户提出的问题是某个应用不可用，通过 IT 服务的管理流程转到后台处理后，发现是由于某个系统软件不可用引起的，通过重新启动该系统软件将不可用应用变得可用，从而解决了某应用不可用的问题。系统软件就是组织的“资源”。

服务战略阶段包括财务管理、服务组合管理、需求管理几个关键过程。

财务管理是从成本效益角度思考问题的，即组织在制定 IT 服务发展战略时，应当考虑 IT 服务发生的成本、增加的利润或者收入，从经济学的角度定义 IT 服务。

服务组合管理是制定 IT 服务战略的主要方法，根据 IT 服务集中服务的价值进行评估，确定服务的组合，以便组织制定服务管理的策略，包括服务投资策略、服务处理策略、服务退出策略等。服务组合管理与业务组合管理的思维类似，只不过服务组合管理面向组织内部的业务部门，而业务组合管理面向组织的外部市场与客户。

需求管理是服务战略制定中的关键一环，可以分为功能性需求和非功能性需求，非功能性需求包括用户体验、可用性、性能等。需求管理过程对需求进行描述和验证，目标是保证 IT 服务管理过程的完整性。

2. 服务设计阶段

服务设计以服务战略为指导，通过对 IT 服务的综合考虑，为服务转换和服务运营做好各种准备工作。良好的服务设计能够帮助组织降低 TCO（总体拥有成本）、提高服务质量、提升服务一致性、使新的或者变更的服务更易于实施、提升 IT 治理能力、保证服务管理与 IT 过程的有效性、改进信息与决策支持等。

服务设计阶段需要考虑的因素包括服务目录、服务水平、服务能力（所需资源）、可用性、IT 服务连续性、信息安全、服务供应商等方面。

服务目录管理是在 ITILv3 中着重提出的，主要原因是 IT 服务需要面向客户、用户等各参与方，是价值创造与成本消耗的依据。通俗地讲，就是 IT 服务为客户带来什么、多大价值，需要消耗多少人力、财力、物力成本，都需要基于服务进行测算。

如果将服务目录比作饭店的菜单，一方面，菜单（IT 服务目录）可以作为客户选择饭菜、酒水等的参考，另一方面，菜单（IT 服务目录）中的条目也反映了饭店消耗的租金、水、电、煤气、蔬菜、厨师等资源成本情况。

服务水平管理是服务差异化的一种表现，所谓“看人下菜”，需要区别对待不同等级的 IT 服务。服务水平管理过程的目标是保证服务提供商能够按照预先约定的服务水平来交付服务，无论是当前正在使用的 IT 服务还是将要使用的 IT 服务。

能力是 IT 服务与资源的连接点，跨越服务生命周期的全过程，在服务设计阶段起到非常关键的作用，可以将能力分为业务能力、服务能力以及组件能力。能力管理应当做到既能满足业务需求，又能以较为合理的成本实现，也就是说能力管理需要平衡成本与资源、

供给与需求。

能力管理可以将当前的能力与规划要求的能力进行对比，以便帮助服务提供商确定哪一个组件需要升级、何时升级以及升级的成本等。能力管理应当植入服务组合与采购过程中，以保证服务提供商与服务供应商的合作共赢。

可用性是交互式 IT 产品/系统的重要质量指标，指的是产品对用户来说有效、易学、高效、好记、错少以及令人满意的程度，即用户能否用产品完成他的任务，效率如何，主观感受怎样，实际上是用户视角的产品质量，是产品竞争力的核心，可见，可用性管理是非常重要的。

为了实现可用性管理，需要提供对可用性相关问题的管理，关联服务与资源，确保所有可用性目标得以测度与实现。一般可用性测度标准包括：可用百分比、不可用百分比、持续时间、失败频率、失败影响等。可用性管理一般要经过检测、诊断、修补、恢复、复位的过程。

IT 服务连续性管理的目标是保证 IT 技术与服务设施（包括计算机系统、网络、应用、数据仓库、通信、环境、技术支持与服务台）能够在要求的、约定的时间内得以恢复。IT 服务连续性管理聚焦在那些重大的足以称之为灾难的事件（非重大事件缺省由事故管理过程来处理）。可见，IT 服务连续性管理类似于容灾管理。为了保证系统的可靠性，可以根据业务特点进行分级管理，比如分为数据级、平台级、应用级等，也可按地域范围分为同城级（城域）、异地级（广域）等。

信息安全管理过程的目标是使 IT 安全与业务安全保持一致，保证信息安全在所有服务与服务管理活动中得以有效管理。信息安全应当保证信息的可用性、私密性、完整性以及在业务过程中的优先级。与信息安全管理有关的策略包括：使用或误用 IT 资产、接入控制、密码控制、电子邮件、因特网、抗病毒、信息分级、文档分级、远程访问、供应商访问 IT 服务/信息/组件、资产处理等。

供应商管理过程的目标是通过对供应商及其提供服务的管理，达到为业务提供无缝的 IT 服务质量的目的，保证物有所值。按照 IT 服务的重要性，可以将供应商划分为战略型、战术型、操作型、一般商品型。不同级别的供应商对于服务提供商的风险不同，采取的管理措施也不一样。与供应商管理相关的信息包括：供应商分类与合同维保，新供应商评估及合约设置，供应商创建，供应商管理与合约履行，续约及解约等。

3. 服务转换阶段

当服务设计完成后，需要将新的或者变更的服务转换为运营态，同时对失败或者破坏的风险实施有效控制，此阶段的工作称为服务转换。

服务转换的策略包括：尽可能使用现有的过程与系统、对齐服务转换计划与业务需求、建立并维系与利益相关者的关系、提供知识转换与决策支持、准备待发送的包、保证新的或者变更的服务质量等。

按照服务转换的先后顺序，服务转换过程包括转换规划与支持、变更管理、服务资产与配置管理、发布与部署管理、服务验证与测试、评估以及知识管理，共 7 个过程。服务转换过程如图 7-3-6 所示。

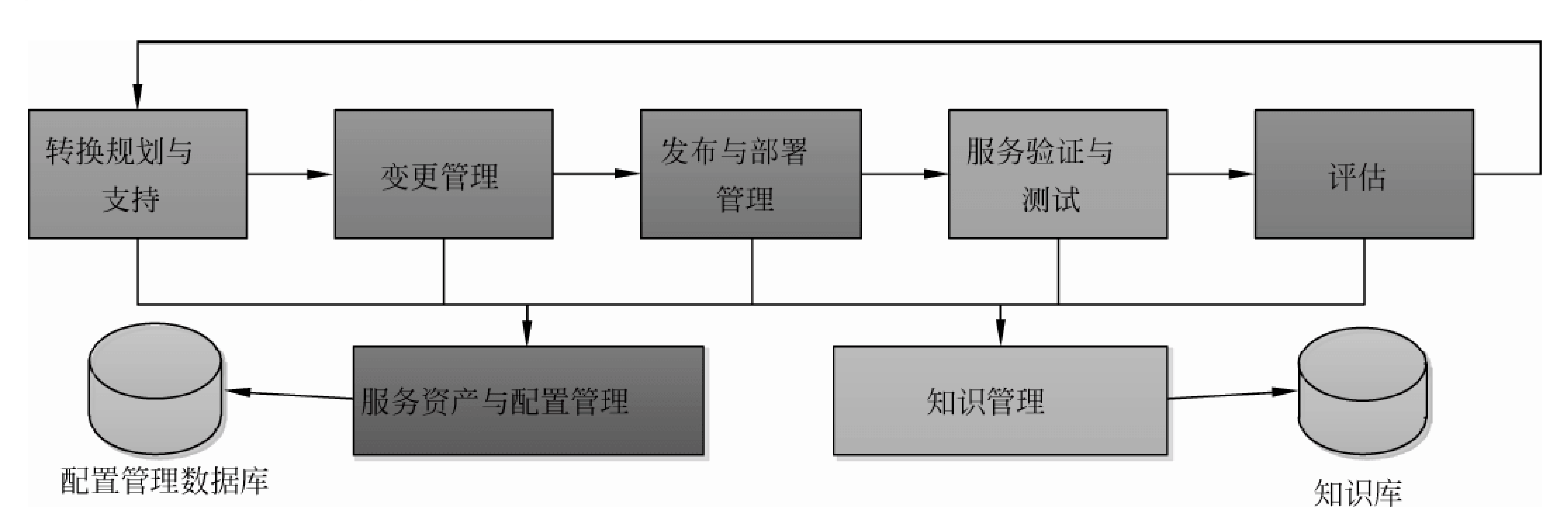


图 7-3-6　服务转换过程

此外，服务转换的过程也可以分为服务生命周期支持与服务转换支持两类。

（1）支持服务生命周期的过程

支持服务生命周期的过程包括：变更管理、服务资产与配置管理、知识管理。

变更是由许多原因引起的，可以归为主动发起的变更和被动处理的变更两类。主动发起的变更是为了寻求业务利益，比如降低成本、提升服务或者提升易用性和支持的有效性。被动处理的变更是作为解决错误并适应变化的环境的一种手段。

变更管理的目的包括：减少风险、降低任何冲击与破坏的严重程度以及提升一次性成功率。根据变更产生的原因，将变更分为战略变更与运营变更两类。

战略变更产生的原因包括：组织变更、法律/规章制度变更、策略与标准变更、分析业务、客户以及用户活动模式后的变更、引入新服务的变更、采购模型引起变化的变更、技

术创新等。

运营变更产生的原因包括服务运营人员需要实施纠错型与预防型变更，因此需要借助标准化的变更过程，例如：服务器重启会影响到共享服务等类似情况。

服务资产与配置管理的目的是辅助完成服务转换过程，包括资产管理和配置管理两个方面。资产管理主要关注资源的价值属性，配置管理则关注于资源的使用属性。资产管理记录了资源所占用的成本，以便于计算 IT 服务所消耗的成本。配置管理通过配置项及其关联关系刻画了应用（人力资源管理、客户关系管理、合作伙伴管理等）、基础设施（服务器、存储、网络等）、机房（电源、空调、机柜、各种传感器等）等资源之间的关系。

配置管理的作用是辅助完成服务转换方案的制定（比如服务变更需要涉及哪些应用、软件、硬件）以及服务部署的执行（根据服务转换方案，将资源配置到相应的软件和硬件设备上）。配置管理的基本单元是配置项（Configuration Item，CI），配置管理功能需要在配置管理数据库（Configuration Management Database，CMDB）的支撑下完成。

知识是数据与信息的高级阶段。如果说数据是离散的事件集合（一般以结构化形式存在），信息来自于由数据支持的上下文（一般以半结构化形式存在），那么知识则是由经验、思想、洞察、价值以及判断组成的，是人类智慧的体现。

知识管理的目的是保证信息能够在恰当的时间、地点交付到能够胜任某项工作的人员手中，辅助其做出明智的决策。知识管理对于成果的服务转换体现为：

（1）用户、服务台、支持人员以及供应商能够理解新的或者变更的服务，包括那些与错误有关的知识，以帮助他们在服务管理中做得更好；

（2）帮助人们意识到当前使用的服务并终止先前的版本；

（3）建立与转换相关的、可承受的风险与信心，例如基于测试结果与其他保障结果正确的量度、理解与行动。

可以说，知识管理在 IT 服务生命周期中具有非常重要的作用，建议组织构建一套单独的知识管理系统，利用知识手段更好地支持 IT 服务的管理。

（2）支持服务转换的过程

支持服务转换的过程包括：转换规划与支持、发布与部署管理、服务验证与测试、评估。

转换规划与支持的目标是协调足够的能力与资源，以便能够以可预测的成本、质量与时间，将新的或者变更的服务变换为生产状态。通过规划与协调各种资源，保证服务战略

的需求、服务设计的编码成果能够在服务运营中得以有效地实现。

转换规划与支持的范围包括：将设计与运营需求纳入转换计划，管理并运营转换规划及支撑活动，管理服务转换的进度、变化、问题、风险与偏差，所有服务转换、发布与部署计划的质量回顾，服务转换过程、支撑系统及工具的管理与运营，与客户、用户及利益相关者的沟通、监视并改进服务转换的绩效。

发布与部署管理的目标是建立、测试与交付在服务设计阶段指定的服务，因此可以满足股东的需求并提供预期的目标。进行软件、硬件的规划、设计、建设、配置和测试，为生产环境创建一系列发布组件。按发布规模，将发布分为紧急发布、小规模发布、大规模发布；按发布种类，将发布分为全发布、Delta 发布（仅少量变更）和包发布。

服务验证与测试过程的目的包括：

（1）计划并实施结构化的验证与测试过程，保证为新的或者变更的服务满足客户的业务及利益相关者的需求提供客观证据，包括一致的服务水平；

（2）对服务组件构成、服务结果以及该版本交付的服务能力提供质量保证；

（3）识别、评估与表达整个服务转换中的问题、错误与风险。

评估的目标是正确设置利益相关者的期望，并为变更管理提供有效的、准确的信息，以便确信影响服务能力与引入风险的变更已经完成了转换检查。所有服务变更的真实绩效是服务提供商的重要信息源，客观的评估能够保证期望值是现实的并能够识别出绩效无法满足期望值的诸多原因。

4. 服务运营阶段

服务运营提供了对 IT 的日常运营进行管理的过程。

服务运营的主要目的是通过一系列日常活动和过程的协调执行，为客户和用户提供可管理的、达到既定的服务水平协议的服务。同时，服务运营也需要对服务提供支持并对过程中所必需的技术进行管理。

服务运营分为服务运营过程（Process）和服务运营职能（Function）两部分。服务运营过程包括事件管理、事故管理、问题管理、请求实现、访问管理；服务运营职能包括服务台、技术管理、IT 运维管理、应用管理。

1）服务运营过程

（1）事件（Event）管理过程。

事件是任何可察觉和可识别的、对 IT 基础设施管理或者 IT 服务造成影响和背离的重

要现象。事件是典型的通知，由 IT 服务、配置项或者监控工具创建。

事件管理过程的目标：确保正常运营而进行的对 IT 基础设施中发生的所有事件进行监控的过程，事件管理也负责对例外情况进行侦测并进行必要的升级。有效的服务运营需要对 IT 设施运行状态的及时掌控和任何对服务偏移的识别，这依赖于有效的监控管理系统。事件管理过程用于需要被控制和可自动化的服务管理的各个方面。

事件管理过程的主要活动为：事件发生后，监控系统产生事件通知后，其检查、发现、转换并理解事件通知，事件过滤过程进行第一次关联性分析并过滤掉不需要处理的事件（比如说排重），之后进行事件的重要性分析。

事件分为三类：第一类是信息类，不需要处理，只需要记录。第二类是告警类，提示服务或设备运行状态接近临界点。第三类是故障类，显示服务或设备运行失败或异常，需要启动事故管理过程、问题管理过程或变更管理过程。对于告警类事件需要进一步的关联性分析，并触发相应的反应，反应可以是记录事件、自动反应、告警并人为干预或触发事故管理、问题管理、变更管理过程。事件处理的行动完成之后要进行效果评估，如果处理结果有效则关闭事件。

（2）事故（Incident）管理过程。

事故是指对一项 IT 服务或一项 IT 服务质量减少的非计划中断，事故比事件要严重得多。

事故管理过程的主要目标是根据服务水平协议的要求，在尽可能小地影响客户和用户业务的情况下尽快将服务恢复到“正常状态”。

事故管理过程包括对服务引起中断或可能中断的事件的管理，包括了用户通过服务台或通过从事件监控工具直接提交的事故。事故由技术员报告和记录，但并不是所有的事件都是事故，许多的事件并不与中断相关，而仅是正常的运营指标或一些简单的信息。

（3）问题（Problem）管理过程。

问题是一个或多个不知原因的事件。

问题管理过程的主要目标是预防问题和事故的再次发生，并将未能解决的事故的影响降低到小。与事故管理强调事故恢复的速度不同，问题管理强调的是找出事故产生的根源，从而制定恰当的解决方案或制定防止其再次发生的预防措施。

问题管理过程包括了诊断事故根本原因和确定问题解决方案所需要的活动，通过合适的控制过程，尤其是变更管理和发布管理，确保方案的实施。问题管理还将维护有关问题、应急方案和解决方案的信息，以使组织能够减少事故的数量和影响。就此而言，问题管理

与知识管理，以及诸如经验数据库等工具有着紧密联系。

问题管理过程包括被动管理和主动管理两种类型。被动问题管理一般作为服务运营的一部分来执行，主动问题管理是由服务运营发起的，但通常是由服务改进驱动的。

（4）请求实现（Request Fulfilment）过程。

请求实现过程主要针对“服务请求”类事件，指的是 IT 部门向用户提供的一系列不同种类的一般需求，这些请求通常可以分为两类：一类是低风险、经常发生且成本低的微小变更；比如重置口令、对某个特殊的主机进行额外软件安装的请求等；另一类为信息咨询请求，由于这些请求是经常发生、低风险的，因而需要采取一个单独的过程来进行管理，而不是混杂于正常的事件和变更管理过程，变成一种累赘和障碍。

请求实现过程的主要目标为：

① 对于某些预定义的申请和需求，为用户提供一个渠道来获得这些标准服务；

② 为客户和用户提供服务请求管理过程的服务和程序信息；

③ 获得和交付请求的标准服务组件；

④ 协助处理一般信息、抱怨或者投诉。

（5）访问（Access）管理过程。

访问管理过程是为合适的用户合理地使用服务进行授权，同时限制未授权用户的访问。访问管理也被称为权限管理或者身份管理。

访问管理过程为用户能够使用一项或一组服务进行授权，因而它是对安全和可用性管理过程所定义的策略的执行。

2）服务运营职能

（1）服务台（Service Desk）职能。

只有服务过程并不能产生有效的服务运营，还需要稳固的 IT 基础设施和适当能力的人员。为了实现这一目标，服务运营依赖于熟练的几组服务支持人员，使用不同的管理过程，充分发挥 IT 基础设施的能力来满足业务需求。

服务台是联系用户的主要接触点，服务台提供了用户与 IT 服务部门之间的联系窗口。当有服务中断、服务请求，甚至某些类别的变更请求时，服务台将为用户提供统一的沟通中心并完成各个 IT 组织和过程的协调。

（2）技术管理职能。

技术管理提供所需要的详细的技术技能和资源来支持 IT 基础设施的持续运营。技术管理在 IT 服务的设计、测试、发布和改进中也起着重要的作用。在小的组织里，技术管理职

能可能由单一部门来管理，但更大的组织通常分为许多专业技术部门。

技术管理职能可以分为服务器相关、存储相关、容灾相关、网络相关、安全相关、计费相关等，解决各种技术问题同样需要知识库的支持。

（3）IT 运营管理职能。

IT 运营管理是根据服务设计过程中定义的性能标准，执行被管 IT 基础设施所需的日常运营活动。在一些组织中这一职能由一个部门来统一管理，而在某些组织中则是一部分活动和人员集中化管理，其余部分由各个专业部门提供。

IT 运营管理分为两个专门的职能。一个是 IT 运营的控制，一般是由运营团队确保日常业务工作的开展。另一个是机房环境管理，负责数据中心物理环境的管理。

（4）应用管理职能。

应用管理负责管理应用程序的整个生命周期。应用程序管理职能支持和维护应用程序的运行并在应用的设计、测试和改进中也起着重要的作用，应用管理通常根据组织的应用组合成若干部门，从而可以提供更专业化、更集中的技术支持。

应用生命周期包括需求、设计、构建、部署、运行、优化几个阶段。

5. 持续服务改进阶段

服务改进简称为 CSI（Continuous Service Improvement），其主要目的是对支持业务过程的 IT 服务进行识别与改进，持续地拉近 IT 服务与变化的业务需求之间的距离。

IT 服务持续改进需要遵循科学的方法，其改进模型如图 7-3-7 所示。

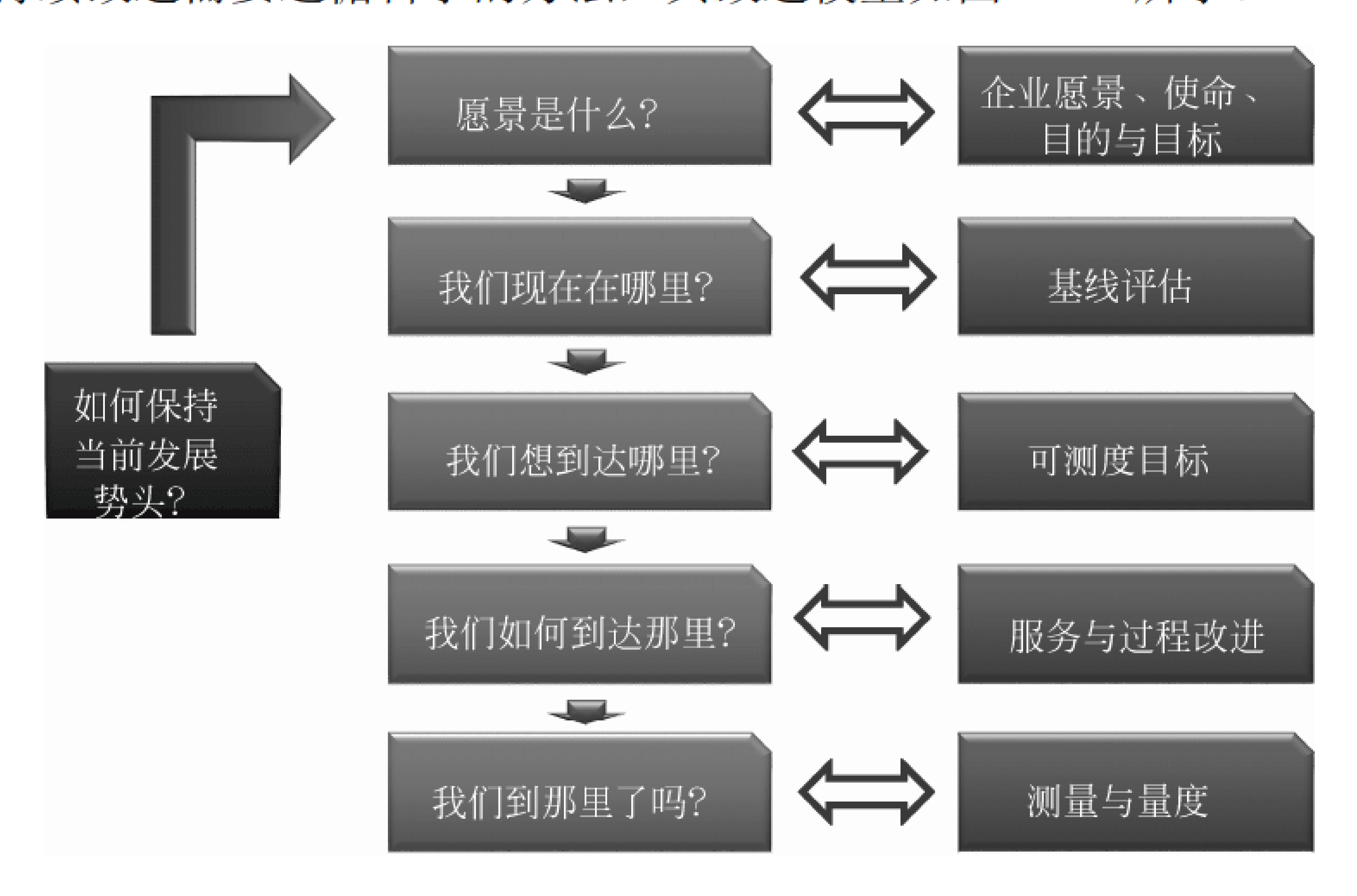


图 7-3-7　服务持续改进模型

IT 服务持续改进还需要遵循科学方法与步骤，逐步实现对 IT 服务的改进。实现 IT 服务持续改进可以参考七步法，如图 7-3-8 所示。

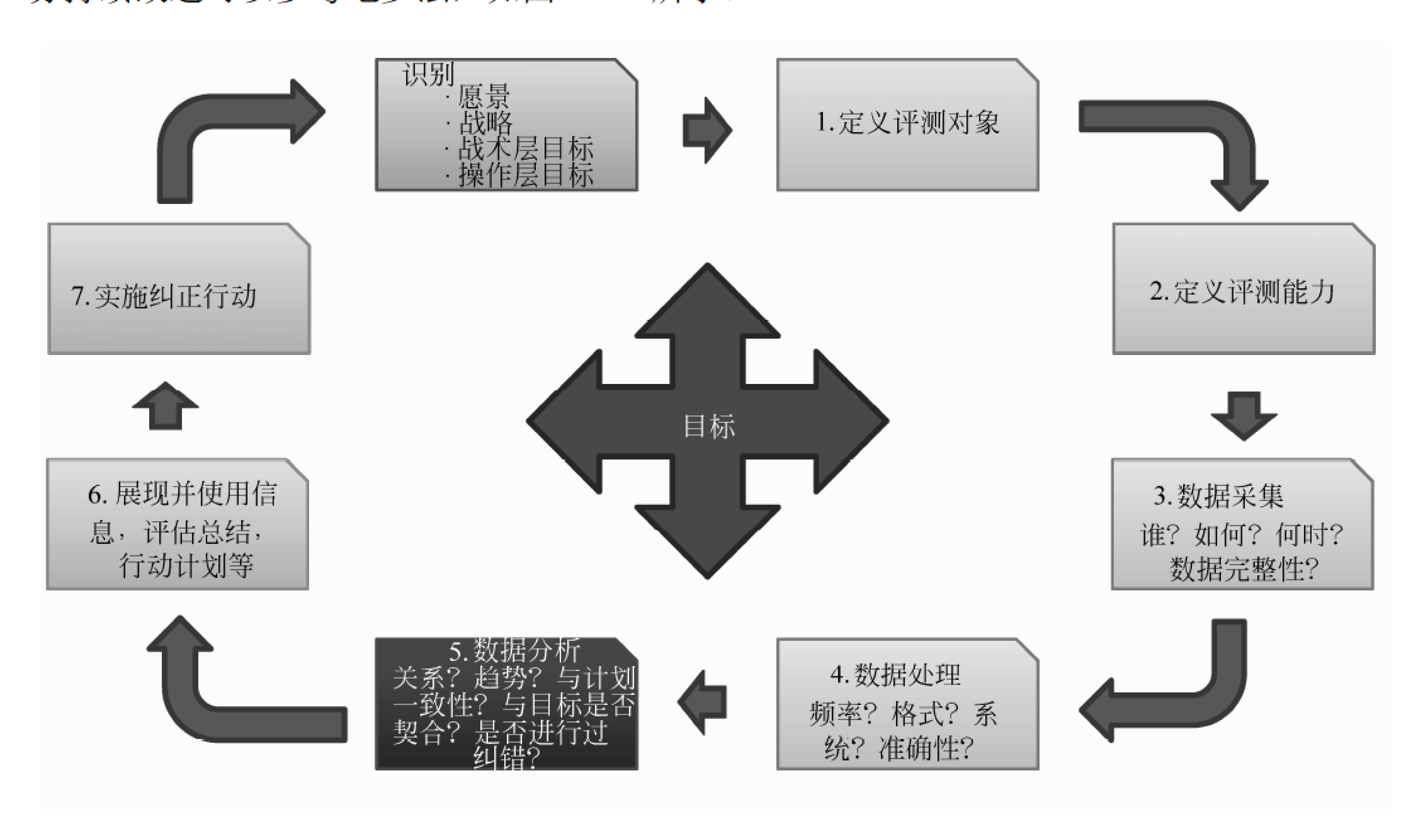


图 7-3-8　IT 服务持续改进七步法

知识管理在 CSI 中起到重要作用。在服务生命周期的每个阶段，数据应当被捕获并形成知识，由此理解实际发生了什么，通过长期的积累沉淀，最终形成智慧，并指导组织对 IT 服务的管理。

7.4　主要内容回顾

企业架构是企业发展战略和企业日常运营之间的桥梁和纽带，如果企业不进行架构设计，则难以将企业发展战略有效地贯彻到日常运营/IT 运维或者之中，难以解决复杂的企业管理问题。

Zachman 企业架构框架是企业架构领域的典型代表，通过自顶而下的 5 个层次（业务范围、业务模型、系统模型、技术模型、详细描述）和从左到右的 6 个 W（What、How、Where、When、Who、Why），清晰地定义了 IT 系统。

除了 Zachman 企业架构模型，开放组架构框架（TOGAF）、集成式架构框架（IAF）、美国首席信息官协会（NASCIO）也是企业架构领域的重要框架体系，以上企业架构模型虽然在设计思路上与 Zachman 模型有一些区别，但是实现原理和方法上基本是相似的。

为保证本书思路清晰、结构严谨，需要以经过长期实践检验的方法论作为指导。Frameworx 和 ITIL 分别为电信行业和 IT 领域的国际最佳实践，可以为本书的编写提供参考。

Frameworx 是四位一体的框架体系，Frameworx 框架体系分为：业务过程框架、信息框架、应用框架、集成框架，共四个维度。其中，集成框架负责将业务过程框架、信息框架、应用框架连接起来。

大数据的核心目标是要解决组织的业务问题，因此大数据运营的本质是将大数据服务植入业务活动的决策环节中，而 Frameworx 的业务过程框架和信息框架恰恰是从动态和静态两个角度，实现了对企业业务活动的刻画，大数据服务可以借助这两个框架体系，实现企业的大数据服务在业务层面的落地。

按照 Frameworx 业务过程框架的分域管理方法，大数据服务可以植入企业的战略、战术、执行三个层面的业务活动之中。

在企业发展战略层面，大数据服务可以辅助企业完成环境分析，比如政治与法律环境分析、经济环境分析、社会与文化环境分析、技术发展趋势分析、竞争对手分析、企业内部资源分析等，辅助企业做出发展战略决策。

在企业的战术管理层面，大数据服务可以帮助企业中层管理人员更好地完成生产与运营决策，比如面向市场部门的广告投放效果分析、面向销售部门的渠道投资效益分析、面向客户服务部门的服务效率分析、面向人力资源部门的人才引进效果分析，等等。

在企业的落地执行层面，大数据服务可以帮助企业基层人员完成客户信用评估、客户偏好分析、热点分析等任务，从而提高企业的整体运营效率，降低了运营风险，提升了客户感知，间接地促进了企业增收。

大数据服务通常没有明确的需求，Frameworx 框架体系中的业务过程框架和信息框架可以作为大数据服务需求分析的起点，实现大数据与企业架构的“联姻”，同时也解决了企业对大数据服务的管理问题。

虽然解决了企业大数据服务与业务活动的结合问题，但是毕竟大数据服务是一种 IT 服务，因此只是到业务活动与大数据服务“联姻”这个层面是不够的，还需要将大数据服

务落地。

ITIL/ITSM 作为以 IT 服务为管理对象的国际最佳实践，以服务战略为指导，从服务设计、服务转换、服务运营到服务持续优化，闭环式的 IT 服务管理体系，体现了软件工程中瀑布式与循环迭代的设计思想，可以作为大数据服务落地实施的方法论指导。

大数据服务与满足企业日常生产经营的操作型应用不同，大数据服务属于分析型应用，大数据服务的主要用途是为企业提供不同层面的决策支持，因此，大数据服务在服务战略、服务设计、服务转换、服务运营以及服务持续优化的各个环节、不同阶段的关注点是不同的，需要企业根据大数据服务的特点，参考 ITIL/ITSM 的管理模式，进行重新调整和适配。

第 8 章
大数据技术：他山之石，可以攻玉

科学技术是第一生产力，尤其是近年来，随着信息技术和互联网的飞速发展，科学技术大大地改变了人类生产和生活的方式。

如果说信息技术提高了人类生产和生活的效率，那么互联网则将人类生产和生活从实体空间拓展到了虚拟空间。在互联网虚拟空间里，人们借助互联网来满足信息获取、知识分享、沟通、交往、购物等需求。人们在浩瀚无边的互联网里遨游的时候，互联网则不断编织着越来越丰富的信息，并记录下了越来越多人类的行为：网页浏览、内容搜索、观点评论、商品交易、咨询建议，等等。

为了系统化地掌握大数据技术并理解大数据技术与企业运营的关系，本章以大数据处理流程为主线，将大数据技术分为 3 类，即大数据采集与存储相关技术、大数据建模与分析相关技术、大数据展示与开放相关技术。

大数据采集与存储相关技术主要分为分布式数据库技术、关系型数据库技术两小类；大数据建模与分析相关技术主要分为分布式计算技术和关系型数据计算技术（SQL）两小类；大数据展示与分享相关技术分为 4 类，即 Web 展示和开放技术、移动客户端展示技术、GIS 展示技术、统计报表展示技术。

新的技术和工具不断出现，层出不穷，作为信息技术和互联网行业的从业人员，既需要拥抱变化，又需要掌握技术和工具背后的原理和方法，提高对大数据技术的理解能力，以不变应万变。

本章在讲述各种技术特点的同时，希望为读者构建一个大数据技术体系框架，掌握大数据技术的适用范围以及企业大数据运营对于大数据技术的要求，将大数据技术相关理论与实际有机地结合起来。

本章内容思维导图如下：

- 第 1节 开源框架Hadoop
  - 介绍HDFS/HBase、MapReduce、Storm等组成的开源生态系统
- 第 2节 大数据存储技术
  - 分布式集群存储技术
    - 从主机与磁盘分离模式转变为计算与存储一体化模式
    - 存取单位从以字节为单位转变为以大数据块为单位
    - 数据存取模式从以行为单位的存取模式变为按列存取模式
  - 关系型数据库存储技术
    - 按时间、地区等维度进行分区以及构建索引是提高存取效率的重要手段
    - 关闭日志记录开关等可以提高数据操作效率，多用插入操作少用更新操作
- 第 3节 大数据分析技术
  - 大数据离线分析技术
    - MapReduce是基于分布式文件系统的主流统计计算模式
    - Map负责提取待统计的列数据，Reduce则负责完成聚合运算，即先微分后积分
  - 关系型数据库分析技术
    - 提高统计效率的前提条件是数据已按计划分区并且有序存放
    - 从长远规划数据仓库，需要要尽量存放细粒度原始数据，提高分析的灵活性
    - 多维模型适用于面向不同主题的个性化分析，可以提高分析效率和易用性
  - 大数据实时流式分析技术
    - 不同于离线计算的批量处理模式，实时流式计算采用单个数据的实时累加
    - 实时流式计算采用将数据主动“喷洒（spout）到”计算“节点的处理模式
  - 数据统计分析技术
    - R语言擅长数据建模、统计与可视化，可以与大数据管理工具结合起来使用
- 第 4节 大数据展示技术
  - Web展示技术
    - JEE是支持Web展示的优秀架构，轻量级框架Struts/Spring/Hibernate已广泛应用
    - 采用Flash控件可以快速实现图形化统计功能，例如FusionCharts、eCharts等
  - GIS展示技术
    - 基于地理信息的空间展示更加形象直观，可用于企业营销渠道、通信网络等方面的分析
    - 矢量法与栅格法分别适合对离散数据和连续数据建模，可根据建模对象的特点选用
    - 基于位置的服务是GIS与定位技术相结合的产物，可以用于推送商品促销信息、导航等
  - 移动客户端展示技术
    - 安卓系统赢在开源开放，IOS系统重在用户体验，技术路线与市场定位相匹配
    - 安卓SDK提供丰富的支持库和良好的应用框架，具有强大的大数据服务展示能力

8.1 开源框架 Hadoop

是一个基于分布式文件系统 HDFS 的框架体系，包括离线计算引擎 MapReduce、实时计算引擎 Storm、内存计算引擎 Spark 等。

互联网每时每刻都在源源不断地产生新的信息和数据，而这些信息和数据大多以半结构化和非结构化形式存在，比如网页、邮件等半结构化数据以及图片、语音、视频等非结构化数据。这些不同媒体形式的、海量的数据难以用传统的关系型数据库来承载，需要新的技术、工具和方法。

为了解决这一问题，谷歌、雅虎、亚马逊、阿里巴巴等领先的互联网公司提出了许多大规模分布式计算和存储技术，这里面以谷歌公司发明的 GFS、MapReduce、BigTable 技术最为典型。

谷歌文件系统（Google File System，GFS）是一种可扩展的分布式文件系统，其主要特点是存储文件容量大、便于扩展并且具有良好的容错性，BigTable 构建在 GFS 之上，是一个压缩的、高性能的、私有的数据存储系统，MapReduce 则相当于 GFS 的引擎，将海量的、不同媒介形式的数据进行切分（Map），以大数据块等形式存入数据库集群之中，并根据统计需要对不同节点上的数据进行聚合（Reduce）处理。

受 GFS 等技术的启发，业界产生了许多类似的大规模数据存取技术和工具，并陆续加入开源组织 Apache（阿帕奇）的大家庭，比如 HDFS、Hadoop MapReduce、HBase、Pig、Hive、Sqoop、Storm、Spark 等。HDFS 即 Hadoop File System，HDFS 的实现原理与 GFS 类似。Hadoop MapReduce 与谷歌的 MapReduce 类似。HBase 是 NoSQL 数据库，采用了列式数据存取模式，与 GFS 的 BigTable 类似，Storm 和 Spark 则解决了海量数据流式计算的问题。

为了解决大数据的管理问题，出现了多种技术框架，为了促进软件技术的发展，出现了许多开源的技术框架，最为典型的就是阿帕奇的 Hadoop 开源项目。

阿帕奇 Hadoop 开源项目非常多，一个简单的开源项目框架体系如图 8-1-1 所示。

从图 8-1-1 可以看出，大数据开源框架可以划分为 3 个域：非实时离线计算域、实时流式计算域、管理域。

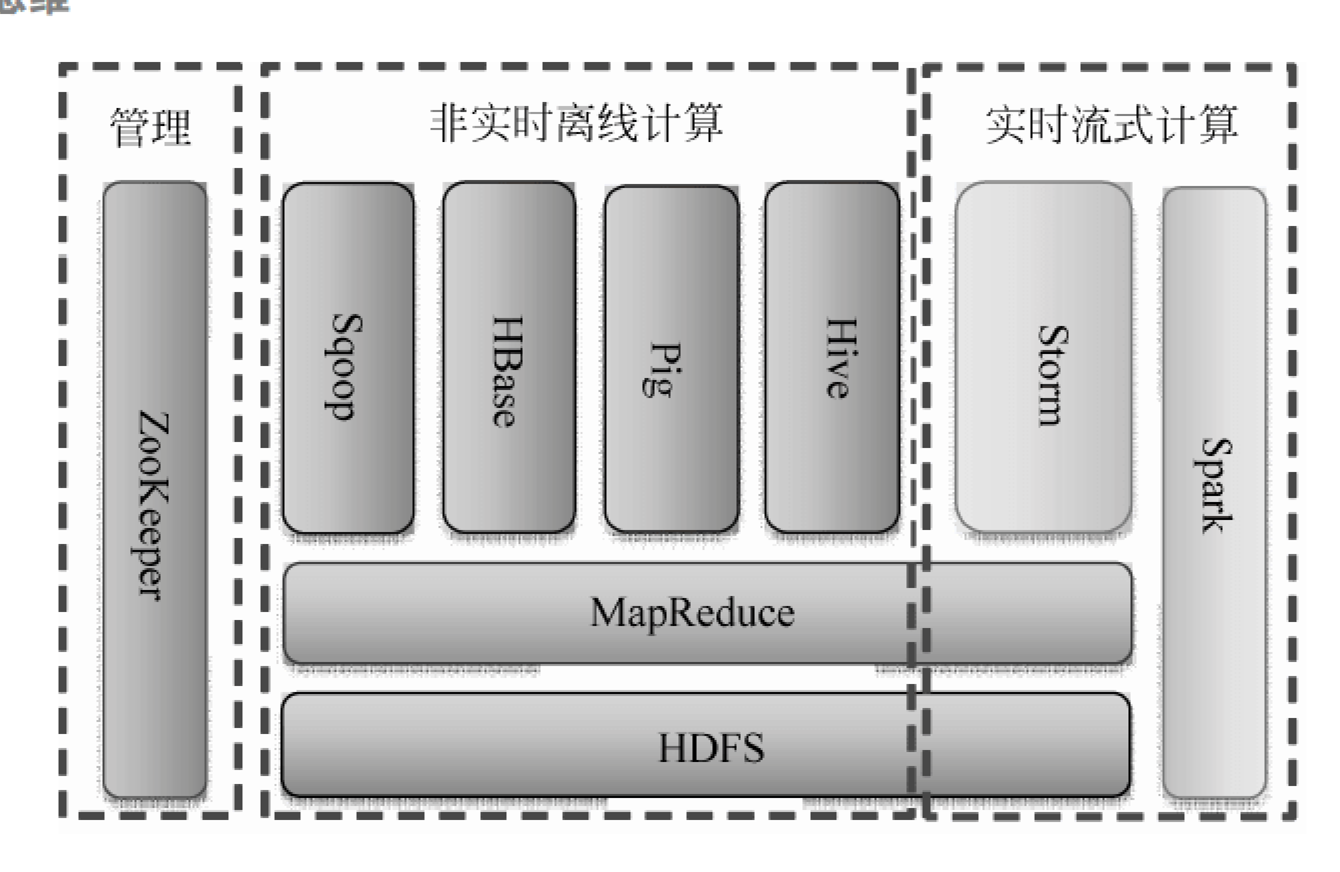


图 8-1-1　一个简单的 Hadoop 开源框架体系

1. 分布式文件系统

HDFS，即 HaDoop File System，是一款典型的开源文件系统，HDFS 位于非实时离线计算的底层，是分布式数据库的基础。

HDFS 与 Windows 操作系统中的 FAT32、NTFS，Linux 操作系统的 EXT3、EXT4 等文件系统相比，是一种面向大文件的文件系统。FAT32、NTFS 等文件系统的数据存取单位为字节，而 HDFS 的数据存取单位通常是一个数据块（典型大小为 64MB）。HDFS 采用以数据块为存取单位的方式，可以大大提高数据的存储容量和存取效率，特别适合对大规模数据的离线处理。

在文件的存取方法方面，HDFS 采用 NameNode 存放文件位置信息，NameNode 类似于操作系统上的目录和文件名，操作系统通过目录和文件名就可以定位文件所在的位置。HDFS 采用 DataNode 存放文件数据。当客户端访问文件时，首先通过 NameNode 来获取文件所在位置，然后根据文件所在位置定位到文件所在的数据节点（DataNode）。NameNode 方式与 Linux 的文件管理方式类似，Linux 借助虚拟文件系统（Virtual File System，VFS）屏蔽了文件操作细节，用户在文件操作时，无须了解被操作文件是一台打印机还是一个数据文件，也无须了解文件实际的部署位置。

当然，为了保证数据的可靠性，Hadoop 会在集群中设置多个副本，这样当主节点或者数据节点出现故障后，就可以重启任务，并将数据访问路径切换到备用节点，保证数据不会丢失。

当 HDFS 中存入大量的数据后，需要借助 MapReduce 完成分析工作。Map 就是按照统

计分析要求，提取数据文件中的统计维度列和统计值列数据，由于原始数据中统计维度列和统计值列是映射的关系，因此称为 Map，Map 就是“映射”的意思。执行 Map 操作后，需要从统计维度列角度对统计值列数据进行排序（Sort），最后再通过 Reduce（聚合）完成统计维度数据项的计算工作，计算动作可以是次数（count）、均值（average）、求和（sum）等。

2. 数据交换工具 Sqoop

Sqoop 是一款位于 Hadoop 和传统关系型数据库之间的数据交换工具，通过 Sqoop，可以实现 Hadoop 与 Oracle、MySQL 等关系型数据库之间数据的导入和导出。负责数据获取的开源框架和工具包括 Pig、Hive 等。

3. 分布式列式数据库 HBase

HBase 架构在 Hadoop 之上，负责大数据的存储。不同于传统关系型数据库，HBase 采用 rowkey 的方式存取数据，数据定义和操作语言采用 NoSQL（Not only SQL），因此又称为 NoSQL 数据库。NoSQL 数据库还包括 BigTable、MongoDB 等。

4. 大数据操作工具 Pig

Pig 是一种针对 Hadoop 数据库进行操作的工具，其实现语言为 Pig Latin，如果没有 Pig，用户需要编写大量的 Java 代码，有了 Pig 工具，用户可以像使用 SQL 那样存取数据。Pig 主要面向大数据应用开发者。

5. 大数据操作工具 Hive

Hive 是一种比 Pig 更方便的大数据操作工具，由于其实现方式与 SQL 非常接近，因此 Hive 的实现语言称为 HiveSQL。

6. Storm

Hadoop 主要适用于大批量离线数据的存取，数据处理的实时性差，而像商品实时推荐、实时风险控制、实时统计等应用对于系统的实时性要求非常高，Storm 框架的出现解决了这一问题。如果说 MapReduce 模型是“计算”找“数据”，那么 Storm 的 Spout/Bolt 模型则正好相反，它采用“数据”找“计算”的方式提高了数据处理的实时性，Spout 就像一

个水龙头，将数据喷射到不同的数据处理节点（Bolt），来一批数据就处理一次，大大提高了数据统计的效率。

7．Spark

Spark 由 Twitter 公司开发并开源，同样解决了海量数据流式分析的问题。Spark 在实现原理上与 Storm 不同，Storm 是将“小数据块”实时地分发（Spout）给“计算”节点，是“数据”找“计算”的思路，而 Spark 则是首先将数据导入 Spark 集群，然后再通过基于内存的管理方式对数据进行快速扫描，通过迭代算法实现全局 I/O 操作的最小化，达到提升整体处理性能的目的，这与 Hadoop 从“计算”找“数据”的实现思路是类似的。可以说，Spark 与 Storm 的整体实现思路基本上是相反的。

不同于 Spark，Spark Streaming 与 Storm 的实现思路基本一致。Spark Streaming 首先对“小数据块”进行批量汇聚，然后再分发给“计算”节点。

Spark 框架支持的编程语言包括 Scala、Java 和 Python。

8．ZooKeeper

ZooKeeper 负责分布式计算环境的管理，功能包括配置维护、名字服务、分布式同步、组服务等。

从以上分布式数据库相关的开源技术可以看出，开源工具的命名都非常有意思，比如 Pig 是猪的英文，Hive 是蜜蜂的英文，ZooKeeper 则是动物管理员，其他工具的名称则是非常形象的动作，比如 Sqoop 意为猛扑，Storm 为风暴，意味着快速，Spark 为火，意味着朝气和力量。

除了开源框架 Hadoop 家族，要完成一个大数据项目，还需要项目管理软件、代码管理软件等作为支持。

微软的 Project 是一款商业版的项目管理软件，OpenProj 是一款开源的项目管理软件，可以跨不同的操作系统平台，适用于小型工程项目。

代码管理工具包括 Git、SourceSafe、SVN 等。Git 是一款开源、免费的分布式版本控制系统，可以敏捷高效地处理任何规模的项目，可以在开发者角色中定义主要开发者和非主要开发者，非主要开发者将软件补丁发送给主开发者。

SourceSafe 是微软公司的代码管理工具，主要面向微软公司的开发工具，如 Visual Basic、Visual C++等。SVN 是 Subversion 的简称，是一款开源的代码管理与版本控制系统。

8.2 大数据存储技术

大数据借助分布式数据库存储，通过软件算法保证数据可靠性，分布式/列式数据库需要与关系型数据结合起来使用。

互联网中的各种信息系统、物联网中的各种传感器产生的大数据首先需要存储，数据的量非常大，有效的解决方法就是将其存放到一个可以动态扩展的分布式存储系统之中。

分布式存储系统需要借助分布式数据库来实现，分布式数据库重点解决大文件存储、存储设备的动态扩展、数据存储节点的容错以及数据的快速检索问题。

分布式数据库技术分为商业和开源两类，它们都以分布式文件系统为基础。开源分布式文件系统以谷歌的 GFS、阿帕奇的 HDFS 最为典型。此外，Pig、Hive、Sqoop 开源工具和框架，可以实现大数据方便、快速地导入、导出与查询。

分布式数据库技术虽然能够解决大数据的存储管理，但并不意味着传统关系型数据库没有了存在的价值。分布式数据库技术难以实现灵活、快速、复杂的统计分析功能，而这恰恰是传统关系型数据库所擅长的，因此，需要将这两种数据库技术结合起来使用，解决不同应用场景下的问题。

关系型数据库包括 Oracle、DB2、SQL Server、MySQL 等，其数据定义和操作语言都是基于标准 SQL 之上的扩展，比如 Oracle 公司的 PL/SQL 就是一款非常强大的数据管理语言，此外，分区、索引、中间表等存储管理技术和方法也在企业数据管理中起到关键作用，对于提升数据的获取效率起到非常重要的作用。

8.2.1 分布式集群存储技术

俗话说“一个好汉三个帮，一个篱笆三个桩”，对于大量的数据存储、计算和传输需求，需要借助集群方式来实现。

公司组织就是集群的一个例子。公司的全部工作可以由一个人来承担，但是当公司业务量大、事务多时，就需要多个不同职能的部门共同承担。比如公司的市场人员负责产品的销售，财务人员负责财务核算，人力资源负责人员招聘、合同管理等。公司通过专业化

的分工与协作，满足了客户需求，也提高了工作的效率。如果企业中有人离职，企业也不想花更多的成本雇佣更多的员工，通常的做法是让两个人互备，这样当某个人无法完成工作时，另外一个人可以代替完成工作。

IT 系统的集群实现原理与公司的管理相似。IT 服务需要借助集群方式，满足大规模高并发的数据处理需求。集群方式可以发挥 IT 资源的整体优势，通过集中 IT 资源满足用户对于计算、存储以及传输能力的需求。

IT 系统集群的实现目标主要包括 5 个方面，即可靠性、可伸缩性、可用性、高性能以及安全性。采用集群方式，不会因为某些 IT 设备或者软件出现故障而导致 IT 服务不可用，集群中任何 IT 设备节点都可能是主节点和备用节点，都可能存储其他节点的备份数据，从而保障 IT 服务的可靠运转。

大数据的特征之一是数据规模大，因此要求 IT 系统能够具有海量数据存储能力，同时数据规模大也意味着需要 IT 系统提供强大的数据处理能力和网络传输能力，而集群方式则可以满足大数据的这些需求。

分布式集群存储技术通常采用以大数据块为单位，将数据切割存储在多个节点中，解决大规模数据存储的问题。为了保证数据的可靠性，通常需要在不同的存储节点中保存多个数据副本。将数据存放到多个节点最大的问题是如何保障数据的一致性，即单次数据操作要求要么成功提交（COMMIT），要么失败回滚（ROLLBACK），不能有中间状态。

为了既能够保证对海量数据的存储，又能够保证事务的一致性，通常对增加、删除、修改、查询操作进行区分处理。多表之间的关联操作是分布式数据库设计的难点。

“增加”操作通常可以采用追加（Append）的方式操作数据，一般比较容易保证事务的一致性。“删除”操作可以采用先标记然后再定期批量删除的方法，这样既能够保证删除的效率，又能够保证及时释放存储空间。

“更新”操作最为复杂，为了保证事务的一致性，通常需要对更新操作先做“插入”再做“删除”，由于整个数据更新过程需要记录操作日志，以便回滚或者提交，因此“更新”操作会消耗大量的存储资源，操作效率低并且容易出错。如果对“更新”操作单独识别和处理，将能够最大限度地保证“更新”操作的效率和成功率。阿里巴巴的开源数据库 OceanBase 就是就是采用单独部署数据“更新”服务器的方式，解决了海量的商品收藏管理问题。

“查询”操作最为简单，因为查询操作只“读”不“写”，无须记录操作状态。目前 Hadoop/HBase 类似的各种列式数据库可以快速装载海量数据，并且可以线性扩展存储容量，实现数据高效率的查询。

在数据库前面增加数据路由层是解决分布式数据库的一种有效方法，数据路由层根据客户端数据 SQL 请求，查询数据库集群中节点的状态，然后将数据操作请求转发到相应的节点，待处理完毕后再将数据处理结果合并起来，反馈给客户端。

在分布式数据库设计时，可以根据应用的特点，采用分别处理增加、删除、修改、查询操作的方式进行架构设计，既要保证事务操作的一致性，又要满足海量数据存取的性能要求。

8.2.2　关系型数据库存储技术

1．关系型数据库的产生

在没有计算机之前，人们通过甲骨、纸等传统媒介来记录和描述人们对于世界的认识，自从有了计算机，信息就成了描述和记载世界的新的媒介方式。

信息通常以文件形式存储和过来，但这种方式使得各种各样的数据变得非常分散，难以很好地关联起来，因此，需要将信息进行归类整理，以结构化、模型化的方式进行归类，这样就可以快速地实现数据的存放和检索。

以上过程其实就是将信息转换为数据的过程，信息是用自然语言描述的、零散的、冗余的，而数据则是计算机能够接受的、严谨的、结构化的。1970 年，IBM 公司高级研究员埃德加·考特（Edgar Frank Codd）提出的《大型共享数据库数据的关系模型》，成为推动关系数据库发展的重要里程碑。后来，关系型数据库飞速发展并得到广泛应用。关系型数据库以关系代数为基础，成为存放和管理数据的有效手段。

结构化数据的主要特征是结构化数据包括的任何一列数据不可再细分，并且任何一列数据都具有相同的数据类型。基于关系代数理论的数据库以数据表为基础，包括并、差、投影、笛卡儿积、选择 5 种基本运算，对结构化数据进行不同方式的关联，可以满足数据维护和统计分析的需求。

2．关系型数据库数据管理语言

当前，典型的关系型数据库包括 Oracle、SQL Server、DB2、Informix、Sysbase、MySQL 等。

关系型数据库的操作语言是结构化查询语言（Structured Query Language，SQL）。SQL

包括数据定义语言（Data Definiation Language，DDL）和数据操作语言（Data Manipulation Language，DML）两部分。DDL 完成数据对象和操作过程的定义，包括数据表、视图、存储过程、触发器、主键、外键、索引、分区等，DML 则完成数据的操作功能，包括增加（Insert）、删除（Delete）、修改（Update）、查询（Select），就是人们经常听到的数据 CRUD（Create、Read、Update、Delete）。

SQL 首先成为数据库语言的美国标准，后来又成为数据库语言的国际标准。为了增强数据库管理能力，数据库软件提供商均在标准 SQL 的基础上进行了扩展，其中以 Oracle 公司的 PL/SQL 最为典型。

3. 关系型数据库数据管理方法

易用性、功能全面性、高性能、安全性、可伸缩性等是评价数据库管理软件的主要方面。在信息技术发展的早期，由于数据规模相对较小并且以事务型应用为主，因此数据库管理软件重点解决高并发条件下的系统响应性能问题。例如，我国的电信、金融、互联网行业中的大型企业，信息系统的用户规模通常在亿级，要应付如此高并发的请求，对数据库管理软件是一个非常大的挑战。

对于大规模用户的高并发请求，一方面需要组织有足够多的 IT 资源作为后台支撑，另一方面还需要具有合理的系统架构，包括存储架构、计算架构、网络传输架构以及容灾架构。

事务型应用通常采用“计算”与“存储”分离的集群架构方式，计算架构和存储架构之间通过光纤网络连接。目前，计算架构和存储架构之间的数据传输完全能够满足要求，数据存取速率瓶颈取决于主机对磁盘数据的存取速率（IO）。

要实现高效地存取数据，首先要解决数据的存放问题，这就好比人们有很多图书，如果图书杂乱无章地摆放，那么快速地找到某一本书几乎是不可能的，如果一本一本地找，那么检索效率也是很低的，如果事先对图书进行分门别类地摆放，就能更快地检索到想要的图书。

与图书分类摆放的思路一样，数据库管理软件采用分区（Partition）、索引（Index）等方式作为存储手段，数据库设计者可以根据查询条件（Where）中字段的使用频率，确定分区或者索引的定义方式。时间段、地域、部门、专业等通常会用作分区的条件，如果数据查询条件为分区条件，系统就可以直接进入某个数据分区查找，而无须进行全库扫描，这种方式大大提高了数据检索效率。

索引的原理是按照索引条件对数据进行重新排序。由于数据是按照从高到低或者从低到高的方式重新组织的，那么就可以采用折半查找法快速定位数据所在位置，而不用一个一个地对比。

在众多关系型数据库之中，甲骨文公司的 Oracle 数据库经过三十余年的发展，以其高稳定性、高性能、易用性等特点，成为世界领先的商业数据库软件。近年来，Oracle 数据库不断顺应信息技术和互联网的发展要求，分别发布了 8i、9i、10g、11g、12c 几个版本。

从 Oracle 数据库的版本号可以看出 Oracle 数据库在不同发展阶段重点解决的问题。

- Oracle 8i 产品是在互联网迅猛发展的 1999 年发布的，其中 i 为互联网（Internet）的首字母。
- 2004 年，Oracle 在分布式计算、网格计算、并行计算等技术发展的背景下，发布了 Oracle 10g 产品，其中 g 为网格计算（Grid Computing）的首字母；2007 年，发布了 Oracle 11g 产品，实现了信息全生命周期管理等多项创新，实现了系统性能、可用性、安全性、开发与测试效率等多个方面的提升。
- 2013 年，甲骨文公司发布面向云计算的 12c 产品，其中 c 为云计算（Cloud Computing）的首字母。

4．关系型数据库在大数据时代的价值和作用

在大数据时代，数据产生的速度和规模都比以往要快、要多，传统的基于关系代数理论的关系型数据库在支持大数据方面显得力不从心。

那么，是否意味着传统的关系型数据库日落西山，逐渐退出历史舞台了呢？当然不是！关系型数据库如同编程语言中的汇编、C、Java 等高级语言的关系一样，是某个特定时代满足特定需求的产物，它并不会随着时代的发展而消亡，只是不太适合新时代数据管理的发展要求，只能专注解决传统领域的问题。

以编程语言为例，汇编、C 等语言并不会随着 C++、Java 等高级语言的出现而消失：C++、Java 等高级语言更容易理解和维护，因此更加适合开发面向用户的应用，而汇编、C 这样的开发语言则适合于处理系统资源占用小、运行速度要求高、更偏向操作系统底层的应用开发。

关系型数据库主要用于支持事务型应用，在面向多用户、高并发的请求的同时，关系型数据库也能够快速地实现数据的增加、更新和删除。事务型应用的特征是交易频率高、交易次数多、单个交易的数据量小。此外，关系型数据库存储的是细粒度的交易型数据，

因此更容易对数据进行排序、分组与合并，可以实现多个维度的数据统计与分析。为了支持 TB 级结构化数据的统计分析，关系型数据库也可以作为数据仓库使用。

与关系型数据库相比，类似 Hadoop/HBase 这样的列式数据库中的数据表结构简单，查询统计需求对数据表之间的关联度要求低，可以满足大规模数据的存储需求，但对于需要多表关联才能完成的复杂的统计分析功能，还需要借助传统的关系型数据库实现，因此分布式集群数据库与传统关系型数据库之间是相互补充的关系。

为了满足大数据时代对于大规模数据的存取需求，同时又能够支持较为复杂的数据查询需求，出现了许多创新型的分布式数据库，这些新型的分布式数据库综合了两种类型数据库的优点，比如 Amazon 的 DynamoDB、Google 的 Megastore、阿里巴巴的 Ocean 开源数据库等。

从新型分布式数据库的实现原理来看，通常是根据不同的应用场景进行了数据操作的创新。以阿里巴巴的 OceanBase 为例，其根据商品收藏等特定需求，区别对待数据的增加、删除、修改、查询操作，对于数据的增加、删除、修改这样的写入操作，通过单独的 UpdateServer 进行管理，而对于查询操作，则通过 ChunkServer 进行管理，通过 MergeServer 完成协议解析、SQL 解析、请求转发、结果合并、多表操作等。

可见，关系型数据库与大规模分布式数据库分别擅长解决不同场景的业务问题，在大数据运营设计时，应当根据两种数据库的特点和优势，制定满足应用需要的数据库解决方案。

8.3 大数据分析技术

大数据典型分析技术为离线计算技术 MapReduce，它以大数据块为操作单位，首先对数据进行微分 Map，然后再对集合内数据进行聚类运算。

分布式数据库和关系型数据库的目标是将大数据存放起来，可是要想在海量数据中发现价值，还需要强大的数据建模和数据分析技术。

大数据建模和数据分析技术与大数据存储技术是不可分割的，不同的数据存储方式决定了不同的数据建模和分析方法。像 GFS、HDFS 这样的分布式数据存储技术将海量数据进行切分并存储到不同的存储节点上，当新的数据产生后，用户无须关心切分后的数据存

放到哪台设备上，数据存储操作对数据管理员来说是透明的，如果存储空间不足，则可以将新的设备添加到集群中。

在分布式存储技术满足日益增长的海量数据存储的同时，也提出了新的问题：如何保证数据获取的效率？如何保障数据的可靠性？如何提高数据获取的便捷性？如何实现分布式数据库与关系型数据库的有效结合？等等。

大数据分析应用分为查询、统计分析、OLAP、数据挖掘几种类型。

大数据查询与传统的交易型数据查询从功能角度看是一样的，区别在于大数据查询解决了海量数据的查询效率问题。

大数据统计分析与大数据查询类似，同样是解决统计效率问题。

OLAP 即在线分析处理，是相对 OLTP（在线事务处理）提出的，OLAP 面向分析，OLTP 面向事务。OLAP 支持多个维度的数据统计。

数据挖掘的目标是从大量的数据中找出看似不相干的事物之间的联系，比如啤酒和尿布之间的联系，某种药物购买行为和流行病之间的联系等。

为支持以上应用，需要有相应的分析技术手段作为支撑。其中，MapReduce 是支持分布式计算的典型分析技术，SQL 是支持关系型计算的典型分析技术。

此外，Storm、Spark 等海量数据实时流式处理技术，弥补了 MapReduce 在海量流式计算方面的不足，R 语言和工具解决了大数据分析结果的展示问题。

8.3.1 大数据建模方法：机器特点与人类诉求

对大数据进行建模的目的是便于对数据进行分析和利用。

以废水处理过程为例，首先是将废水引入，然后再对水进行逐级处理和过滤，最后将处理好的水注入不同的输出管道，比如灌溉渠道、中水渠道、工业用水渠道等，如图 8-3-1 所示。

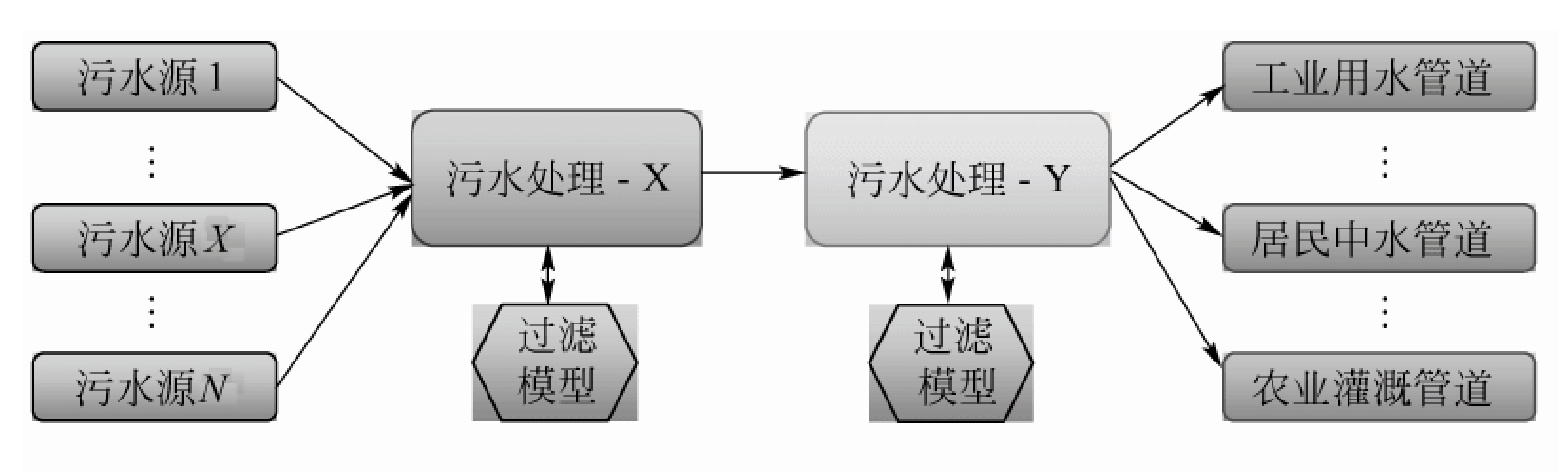


图 8-3-1 废水变可用水的处理过程

数据处理过程与废水处理过程相似。数据处理过程是首先对不同来源的原始数据进行ETL/ELT（加载、转换、清洗），然后将ETL后的数据放入不同的数据模型中，接着根据应用需要对数据进行二次ETL，再将数据放入新的数据模型，最后将数据转入不同的展示渠道或开放渠道，比如桌面Web、移动客户端、短信平台、合作伙伴数据平台等。从原始数据到可用数据的处理过程如图8-3-2所示。

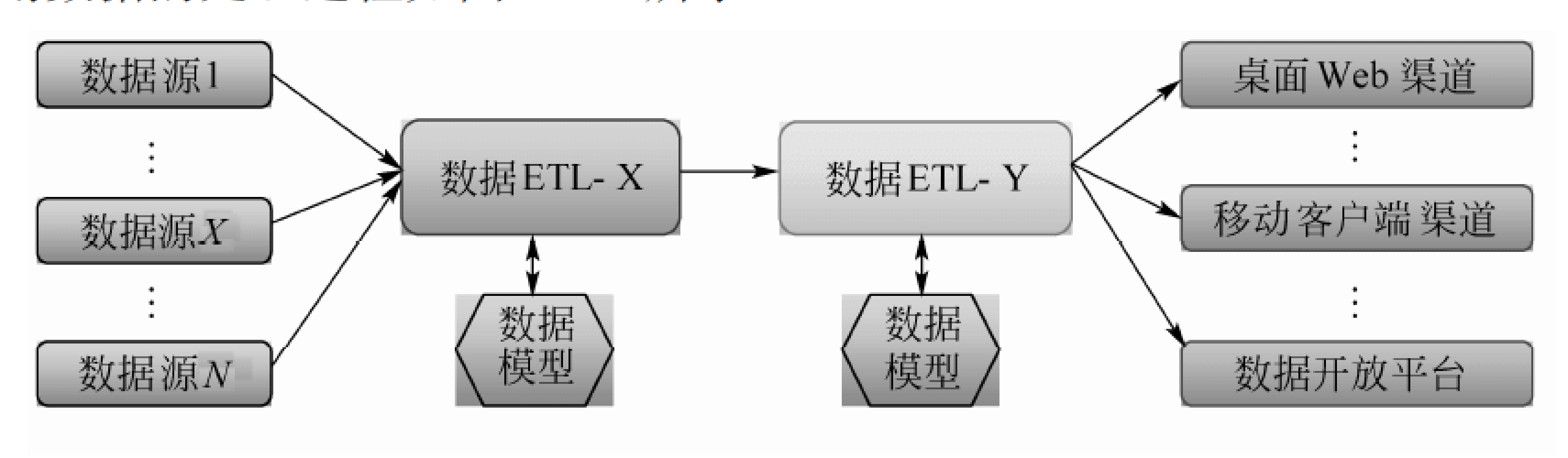


图8-3-2　从原始数据到可用数据的处理过程

从图8-3-2可以看出，数据从原始状态变为可用状态，需要经过多个阶段的ETL，而每个阶段的ETL必然需要数据模型的支持。比如在对多个数据源进行ETL的阶段，需要通过数据模型来装载聚合于不同数据源的数据，如果这些数据还不能满足数据分析应用要求，就需要对数据再次进行ETL，而ETL的基础是数据模型。

数据模型应当支持数据一步步地朝着目标应用靠拢。通常，数据离原始数据越近，数据的颗粒度越小，更便于从多个维度对数据进行分析，同时由于数据颗粒度小，数据处理的时间也较长。另一方面，如果数据离用户越近，那么要求数据的颗粒度要大，数据分析结果要使人容易理解，并能够快速看到数据分析的结果。数据规律如图8-3-3所示。

从图8-3-3可以看出，机器具备的能力和人的需求之间是存在天然鸿沟的：机器侧重于逻辑处理，机器可以通过模型存储海量/单个数据，而人的决策需求则是信息、知识、归纳性、个性化的。



图8-3-3　数据建模过程实际是要不断填平机器与人之间的鸿沟

根据以上分析，对数据从采集到分析的数据建模过程可理解为：

- 在从各种数据源抽取数据并放入数据仓库的初始阶段，尽量通过数据模型放入较小颗粒度的数据，以便从多个维度对数据进行分析。
- 根据用户需要，将数据聚合到颗粒度更大的数据模型中（有的组织称为数据集市）。当数据经过多次 ETL 并放入数据模型后，数据应当变得越来越容易理解和使用，数据的个性化程度越来越高，直至个性化到分析结果能够满足部门、角色、岗位甚至单个特定人的需要。

8.3.2 关系型数据库分析技术

基于关系代数的数据库理论主要解决结构化数据的管理问题，涌现出的数据管理与数据分析理论和实践包括数据仓库、数据集市、操作型数据仓库、分析型数据仓库、在线分析处理、数据挖掘、商业智能等。关系型数据库/数据仓库的分析语言和工具以结构化查询语言（SQL）为主。

关于如何实现大量历史数据的有效管理和分析，数据仓库大师 Bill Immon 和 Ralph Kimball 提出了不同的理论和方法体系。

Bill Immon 强调“自顶而下”的数据仓库构建方法，即先构建一个大的细粒度的数据仓库池，然后再构建面向不同主题的数据集市，这样可以保证数据基础的全面性，基于更细粒度的数据进行分析也会让数据分析更加灵活。

Ralph Kimball 强调“自下而上”的敏捷的数据仓库构建方法，即先构建小的面向不同主题的数据集市，然后再逐步完善数据集市，侧重于借助多维模型的设计，使分析结果更贴近于数据分析用户，有更好的用户体验。

以上两种数据仓库设计大师的观点各有侧重，Bill Immon 的观点侧重于数据仓库的长远规划，而 Ralph Kimball 则侧重于数据仓库对于个性化需求的支持。

在商业领域，有 Oracle、DB2、SQL Server 等企业级数据库，为了支持数据分析应用，设计开发了多种产品和工具。比如甲骨文公司的 Oracle Warehouse Builder，IBM 公司的 DB2 Data Warehouse Edition，微软公司 SQL Server 系列的 Analysis Services Data Mining、Reporting Services 等。

甲骨文公司的 Oracle 数据库是大规模数据管理和分析产品以及工具的集大成者，典型的数据库管理技术和工具包括实时应用集群（Real Application Clusters，RAC）、Partition

（分区）等，Oracle 数据库基于标准 SQL 做了许多扩展，形成了自己的过程化 SQL 语言（Procedural Language/SQL，PL/SQL），PL/SQL 是一种非常强大的 SQL 编程语言。

数据存储方式决定了数据分析的效率，而根据应用特点对数据分区是提高数据查询和分析效率的有效手段。比如，可以从日期、地域、部门等维度对数据进行分区，当数据向数据库装载时，数据库管理系统会根据数据内容将数据分别存放到不同的区域，这样当操作数据时，就会快速地定位数据的存放位置，完成数据的增加、更新、删除、查询等操作。

在数据仓库中预先构建中间表也是提高前台查询和展示效率的一种有效手段。中间表存放汇总后的数据，数据颗粒度更大，数据规模更小，因此能够更加快速地展示分析结果。使用中间表的设计方式主要适用于基于 ODS 的 OLAP 应用。通常采用从多个数据源 SELECT 然后 CREATE 的方式构建中间表，这样可以提高中间表的构建效率，但是这种方式的缺点是无法预先定义分区，仅仅适用于数据量小的应用场景。

如果要创建数据分区，必须要在创建数据表的同时完成分区的创建，预先为数据划分出数据存放空间，完成数据规划。由于在数据操作时，为了保证事务的完整性，通常要记录数据操作日志，为了提高数据导入效率，可以事先取消日志（NOLOGGING）。由于数据操作通常会因为寻找空闲存储空间而降低了数据存取效率，可以在数据操作之间将操作模式改为追加（APPEND）模式，这样就可以将数据在连续的数据空间追加，从而提高数据导入效率。如果涉及多个数据表关联更新，建议采用先 SELECT 再 INSERT 的方式，避免直接使用 UPDATE 操作数据。

构建索引也是提高数据查询和统计效率的有效方式。数据表索引能够提高数据操作效率的原理很简单，就是根据索引预先对数据进行排序，这样就可以通过折半查找法快速找到符合条件的数据，而不是一个一个地对比无序存放的数据。另外，数据表索引也可以存放在不同的物理磁盘，从而提高了数据的并发处理效率。

尽管传统关系型数据库采用单机模式设计，难以无限横向扩展，只能依赖提升单机处理能力的纵向扩展模式，但是关系型数据库在大规模并行计算方面也取得了非常大的性能改进。比如 Oracle 的 g 系列和 c 系列数据库产品，就是采用网格计算、云计算技术研发而成的。即便如此，传统数据库架构受限于采用“集中控制”思维，而“集中化”和“中心化”架构方式不太可能实现系统性能的线性扩展，集群内数据库节点规模非常有限。随着数据库节点的增多，集群总体性能急剧下降。像 Hadoop、HBase 这样的分布式数据库采用了与关系型数据库不同的“去中心化”架构设计思维，处理节点之间是“平等”的，因此可以实现处理能力线性扩展，能够满足大规模数据的存储和分析需要。

8.3.3　大数据离线分析技术

“主机+磁盘阵列”的系统结构实现了计算能力与存储能力的分离，适用于数据存取次数多、单个数据量小的事务型应用。

但是，大数据时代都是海量数据（PB 级）的处理，在面对海量数据存取时，“主机+磁盘阵列”式的系统架构需要在主机集群与磁盘阵列之间消耗大量的网络带宽，这种架构方式大大降低了系统的处理效率，无法满足大数据时代的数据存储要求。

为了解决大规模数据的处理效率问题，需要开启不同的思维模式：能否将数据存储在主机内部，而不是主机外部的磁盘阵列上？如果可以这样，就会大大提高数据的存取效率。是否可以以大数据块（比如 64MB）为单次数据存取单位，而不是以字节为单次数据存取单位？这样可以一次处理大批量数据，进而提高数据的处理效率。

谷歌公司沿着以上思路提出了 GFS、MapReduce、BigTable 这样的大规模分布式数据处理方案，解决了互联网 Web 大规模数据的存储和快速检索问题。GFS 就是以上提到的以大数据块为单位进行数据存储的，而 MapReduce 则相当于基于 GFS 上的数据分析引擎。MapReduce 的数据处理过程如图 8-3-4 所示。

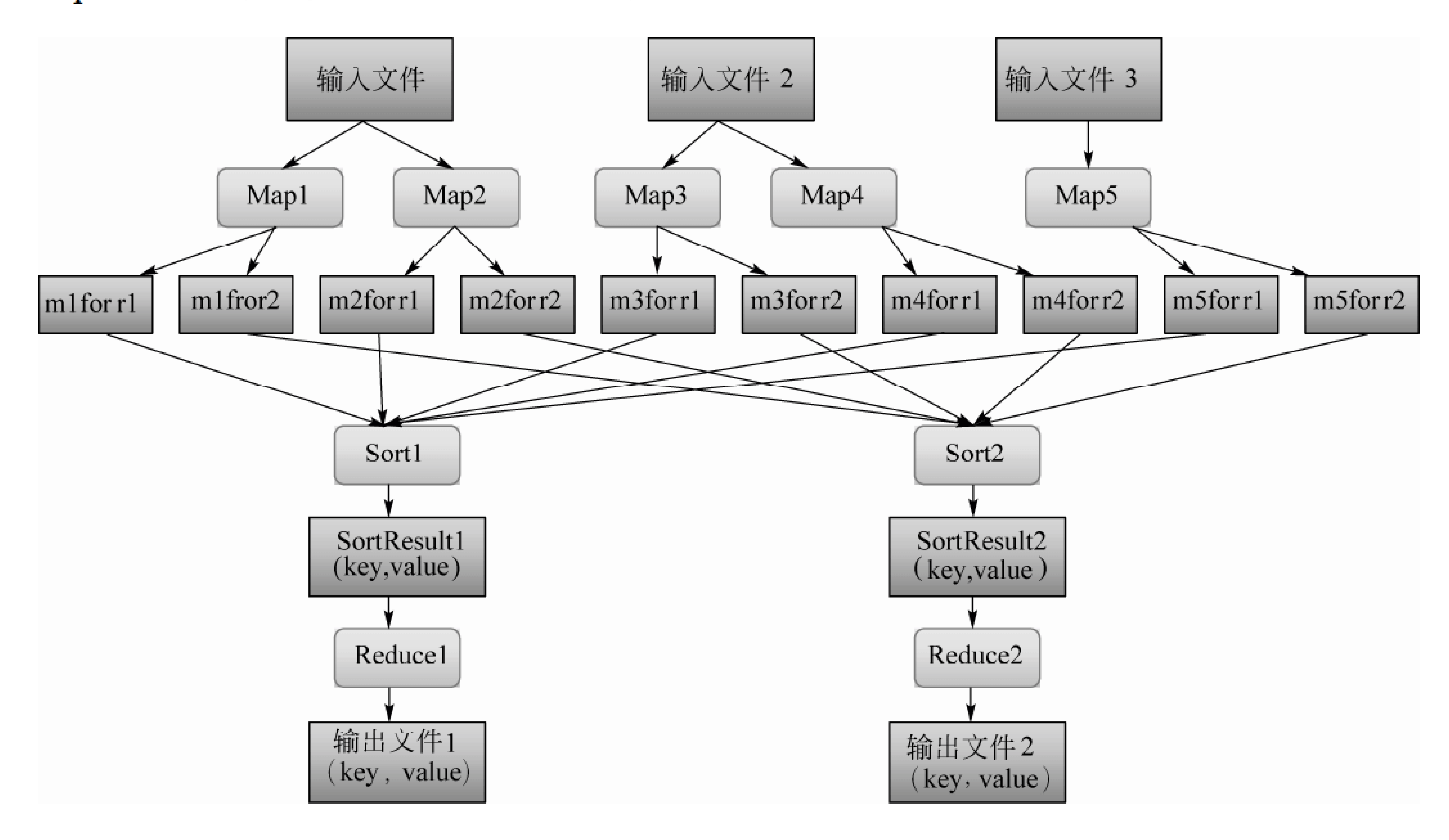


图 8-3-4　MapReduce 数据处理过程

从图 8-3-4 可以看出，输入文件 1 到输入文件 3 为原始数据，原始数据经由 Map、Sort 和 Reduce 3 个子过程，最后输出满足统计要求的结果文件。

为了更加直观地看到 MapReduce 的数据处理过程，下面以电信运营商海量上网记录统计为例，从地域维度统计上网数据流量，上网记录原始数据如下所示。

记录号	上网流量/B	上网日期	手机号码...
第 1 条：	12 345 678	131223	8613812345678...
第 2 条：	56 781 234	131223	8613946571234...
第 3 条：	23 456 789	131223	8613712345678...
第 4 条：	45 678 901	131224	8613546757122...
第 5 条：	33 554 567	131224	861351212111...

以以上 5 条上网记录为原始数据源，按照日期统计上网流量，简单起见，上网记录样本数据中仅保留了本案例所需的上网流量（最左一列 8 位数字）、上网日期（中间一列）和手机号码（最右一列）。

假如以第 1 条和第 2 条上网记录作为输入文件 1 的内容，第 3 条和第 4 条上网记录作为输入文件 2 的内容，第 5 条上网记录是输入文件 3 的内容，如下所示。

第 1 条：	12 345 678.	131223	8613812345678...	输入文件 1
第 2 条：	56 781 234	131223	8613946571234...	
第 3 条：	23 456 789	131223	8613712345678...	输入文件 2
第 4 条：	45 678 901	131224	8613546757122...	
第 5 条：	33 554 567	131224	861351212111...	输入文件 3

3 个输入文件分别首先经过 Map 操作，从输入文件指定位置提取键值对（即本例中的日期为“键”，本例中的“值”为上网流量）到不同的 Map 集合中，作为 Sort 的输入，如下所示。

第 1 条：	**12 345 678**	**131223**	8613812345678...	输入文件 1
第 2 条：	**56 781 234**	**131223**	8613946571234...	
第 3 条：	**23 456 789**	**131223**	8613712345678...	输入文件 2
第 4 条：	**45 678 901**	**131224**	8613546757122...	
第 5 条：	**33 554 567**	**131224**	861351212111...	输入文件 3

提取后的键值对包括 5 个，分别如下：

{131223,12 345 678}、{131223,56 781 234},{131223,23 456 789}、{131224, 45 678 901}、{131224,33 554 567},这些键值对经由 Sort 操作将其放入相应的分组中，分组结果如下所示。

Sort1: {131223,[12 345 678,56 781 234,23 456 789]}

Sort2: {131224,[45 678 901,33 554 567]}

Sort 操作后的结果又成为 Reduce 的输入，其中 Reduce1 的输入为 Sort1,Reduce2 的输入为 Sort2。Reduce 实际上是执行聚合操作，根据预先编制的 Reduce 程序，需要对多个值执行 Sum 操作，因此 Reduce 后的结果为。

Reduce1：{131223,92 583 701}

Reduce2: {131224,79 233 468}

以上就是 MapReduce 的输出结果：2013 年 12 月 23 日的总流量为 92 583 701B（约合 92.58 MB），2013 年 12 月 24 日的总流量为 79 233 468B（约合 79.23 MB）。

以上就是 MapReduce 从多个文件中通过 Map、Sort 和 Reduce 操作进行统计分析的过程。由于以上仅仅是一个剖析原理的简单示例，在 MapReduce 的实际运行环境中，会有大量的输入文件，由于大量的输入文件分布式地存储在不同的主机设备中，并且被分割的大文件无须去主机外部的磁盘阵列上存取，只需在主机内部的磁盘上存取，解决了因大文件传输而引起的大量网络带宽占用问题。由于 Map、Sort、Reduce 操作分别在不同的主机上，通过多个任务并行执行，彼此之间不存在关联依赖，大大提高了数据统计效率。

MapReduce 是一种典型的大数据分析技术，但是也存在许多不足，为了克服这些不足，许多软件在此基础上进行了改进和完善，包括由 Facebook 开发并开源的分布式 NoSQL 数据库软件 Cassandra，开源分布式文件系统 Ceph（支持对象存储），Cleversafe 公司的分散存储网络（将元数据分散到集群中，所以称为分散存储），IBM 公司的通用并行文件系统（General Parallel File System，GPFS），EMC 公司的 Isilon、MapR 文件系统，NetApp 公司的 Hadoop 开放方案等。

8.3.4 大数据实时流式分析技术

在现实生活中，许多应用场景要求系统能够实时做出响应，比如商品实时推介、广告投放、消费额度提醒、实时的风险控制、实时统计、网络故障预防、无线带宽分配、热门话题推送、汽车超速报警等。

系统只有具备良好的实时性，企业才能够把握商机，具备更强的能力。例如，商品实时推介、广告投放等能够提升企业产品销售能力，实时的风险控制、实时统计能够提高企业的管理能力，消费额度提醒可以提升企业的客户服务能力。可见，实时计算可以有效地提升企业的竞争能力。

Hadoop 实现分布式计算的原理是首先借助 HDFS/HBase 将海量数据切片后存入集群中，然后根据数据统计需求，采用 MapReduce 从集群中对数据进行提取（Map）和聚合（Reduce），这种架构方式适合于大规模数据的离线批量处理，但却无法满足实时计算的需求。

为了弥补 Hadoop 在支持大规模数据实时计算方面的不足，解决应用响应的实时性问题，业界提出了许多分布式实时流式计算框架，以 Twitter 开源的 Storm 和 UC Berkeley AMP lab 开源的 Spark 最为典型。

为了清晰地看到分布式实时流式计算技术如何解决大规模数据的实时计算的思路和方法，下面以 Storm 的实现原理为例进行简单分析。

Storm 开源框架包括的概念有 Nimbus、Zookeeper、Supervisor、Worker、Task、Topology、Spout、Bolt、Tuple、Stream、Stream Grouping，它们各自的分工如下。

- Nimbus：主要负责资源分配和任务调度，与 Hadoop 的 JobTracker 相对应。
- Zookeeper：负责维护配置信息、命名服务、分布式同步、分组服务。
- Supervisor：负责接受 Nimbus 分配的任务，启动和停止 worker 进程，与 Hadoop 的 TaskTracker 相对应。
- Task：worker 中执行 spout/bolt 的线程。
- Worker：运行具体处理组件逻辑的进程，worker 中包含一个或者多个 task，与 Hadoop 的 Child 相对应。
- Topology：是反映数据处理的拓扑结构，与 Hadoop 的 Job 相对应。
- Spout：意为“喷射”，就像自来水一样，采集数据源并将其发送到 bolt，与 Hadoop 的 Map 相对应。
- Bolt：接受数据任何执行的组件，执行动作包括过滤、函数操作、合并、写数据库等，与 Hadoop 的 Reduce 相对应。
- Tuple：即元组，一次消息传递的基本单元。
- Stream：多个 tuple 就组成了源源不断的 stream。
- Stream Grouping：流分组策略告诉 topology 如何在两个组件之间发生 tuple。分组策略包括 shuffle、field、all、direct 等。shuffle 是随机发送方式，direct 为指定目的地分组发送方式，field 为按字段分组发送方式，all 为广播发送方式。

为了直观地看到 Storm 的实现原理，从基于 Topology（Spout/Bolt）的流式设计、流分组方式设计、分布式集群设计 3 个层面进行剖析。

基于 Topology（Spout/Bolt）的流式设计思路如图 8-3-5 所示。

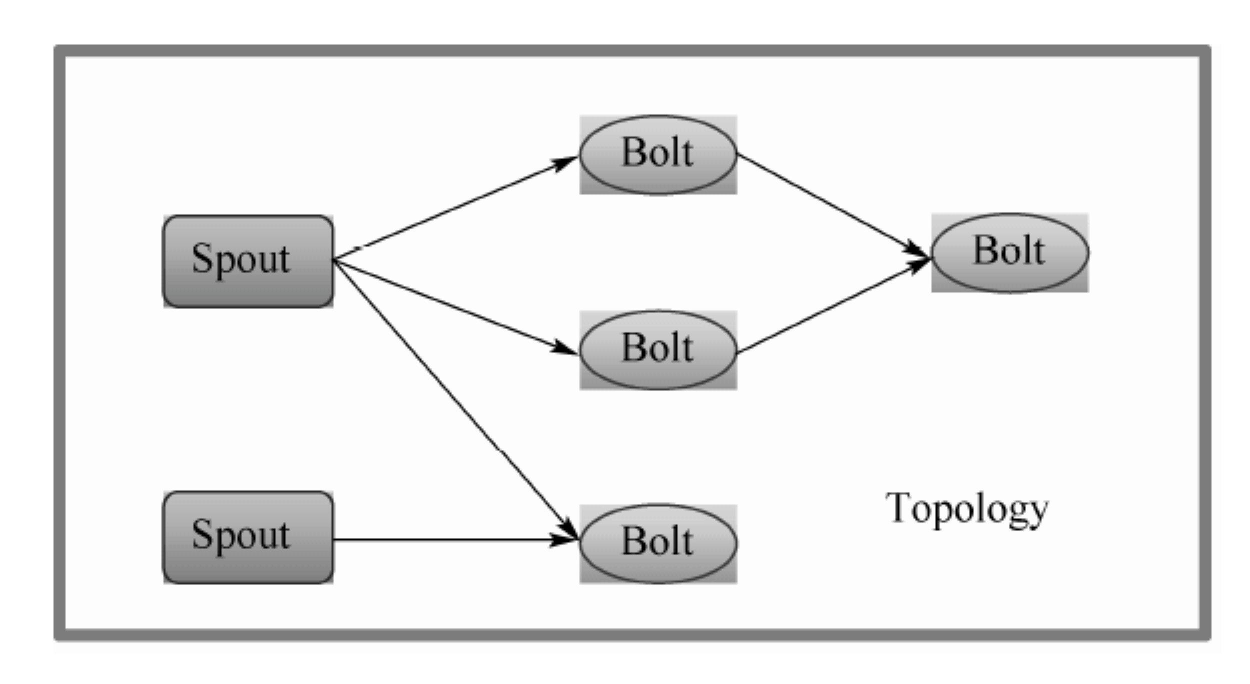


图 8-3-5　Storm 的 Topology 结构

从图 8-3-5 可以看出，Storm 的 Topology 实际上是对 Spout 和 Bolt 之间关系的定义，Spout 是数据源，为 Bolt 注入数据，Bolt 的数据源既可以来自 Spout，也可以来自其他 Bolt。

来自 Spout 的数据流需要发送到 Bolt 进行处理（过滤、汇总、写数据库等），Spout 输出的数据以怎样的方式发送到 Bolt 的则需要通过 Stream Grouping 进行定义（随机发送、按字段发送，还是指定目的 bolt 发送），Stream Grouping 如图 8-3-6 所示。

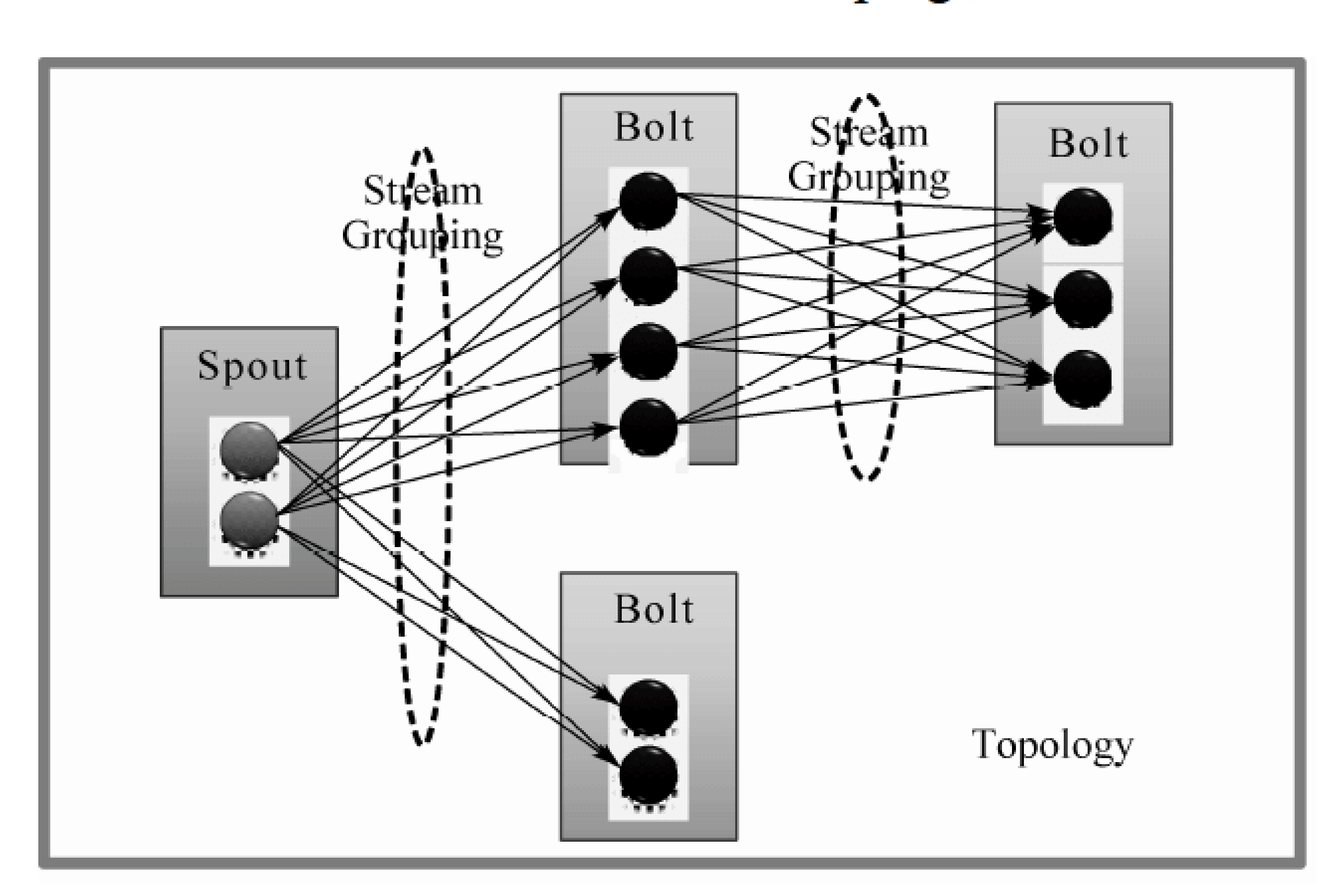


图 8-3-6　Storm 流分组（Stream Grouping）原理示意图

从图 8-3-6 可以看出，可以通过流分组的方式决定将数据流发送给哪个 Bolt 处理，这样就可以按照预先定义来处理数据。比如用户 Web 行为偏好实时（浏览、检索）统计，当用户浏览某个网页或者按照某个关键字搜索后，Spout 就可以将用户的行为数据发送到 Bolt 处理，Bolt 可以按照流分组策略（比如按字段分组）发送到指定的 Bolt 进行统计（总数加 1），这样就可以实时看到用户浏览网页的次数或者搜索关键字的次数，这些统计数据可以

作为热门商品和热门搜索关键字展示给用户。

Storm 框架的优势在于实时处理大规模数据，因此当完成 Storm 应用开发后，需要将其部署到分布式集群中，Storm 的分布式部署架构实现方式如图 8-3-7 所示。

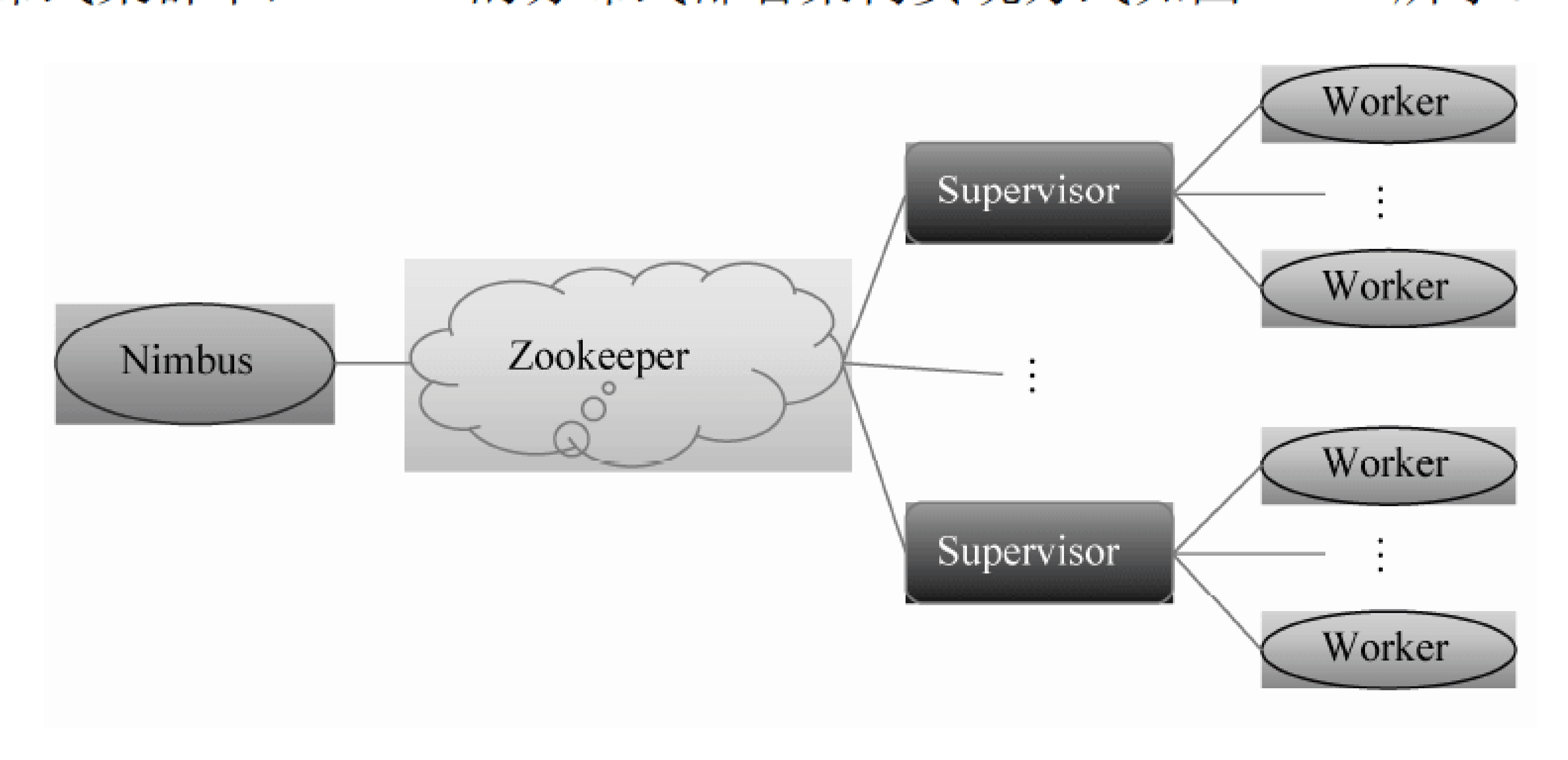


图 8-3-7　Storm 的集群管理方式

从图 8-3-7 可以看出，Storm 采用三级管理方式。第一级是 Nimbus，根据集群资源的占用情况进行资源分配和调度；第二级为 Zookeeper，Zookeeper 为动物管理员的意思，负责维护集群配置信息、分布式同步以及分组等工作；第三级为 Supervisor，Supervisor 负责接受 Nimbus 发来的任务，启动或者停止 Worker 任务。

如果将 Storm 的集群管理模式与企业管理模式对比，那么 Nimbus 则相当于企业的 CEO，在企业全局层面上分配和调度人、财、物等资源；Zookeeper 则相当于企业的分管副总，负责对各个职能部门的工作进行同步，以保证企业能够按步骤、有序地完成任务，而 Supervisor 则相当于各个职能部门的经理，负责传达 CEO 的命令，比如开始干活或者停止干活；Worker 则相当于企业的基层员工，负责根据职能部门经理的要求完成指定的工作。

除了开源框架 Storm 和 Spark 之外，还包括许多分布式实时流式计算技术，例如 Yahoo 的 S4（Simple Scalable Streaming System）、IBM 的 StreamBase、微软公司的 TimeStream，Facebook 的 Data Freeway and Puma 等。

8.3.5　数据统计分析技术

Hadoop/HBase 等 NoSQL 数据库虽然能够通过集群方式存储大规模数据，但是基于大规模数据进行统计分析还需要借助专业的统计分析工具。

用于统计分析的专业软件工具包括 SPSS、SAS、Stata、Excel、Matlab、Amos、R 软件等。

统计产品与服务解决方案（Statistical Product and Service Solutions，SPSS）的特点是简单易用，属于傻瓜式软件，在国内的应用最多。

SAS 则正好相反，通过编程可以实现非常强大的功能，但是比较难掌握，学习周期长，多用于企业之中。

Stata 则介于 SPSS 和 SAS 之间，既有适用于菜鸟的快速上手功能，又具备适用于高手的编程功能，多用于医学、生物统计等方面的研究，在学术界应用广泛。

Excel 是微软公司的产品，非常简单易用，还可以通过宏语言编程实现数据统计功能，并且可以将统计结果直接装载到 Word 之中。

Matlab 能够解决各种各样的数学计算问题，是数学建模的首选工具。

Amos 可以同时处理多个变量，无需编程就可以快速地实现统计分析功能，可以检验数据是否符合所建立的模型以及进行模型探索。

R 软件是一款 GNU 系统的开源软件，其最大的特点是开源和免费，通过 R 语言进行编程，可以实现统计、预测分析以及数据可视化功能。此外，R 软件采用命令行格式，可以集成非常多的数据源。自 2010 年开始，Oracle 数据管理软件开始支持 R 语言，显示了 R 语言在数据分析领域的地位和发展潜力，下面从 R 软件的数据源支持、数据结构以及应用场景 3 个方面进行介绍。

R 软件集成的数据源包括统计软件 SPSS、SAS，文本文件 XML、ASCII，结构化数据库 Oracle、MySQL 等，如图 8-3-8 所示。

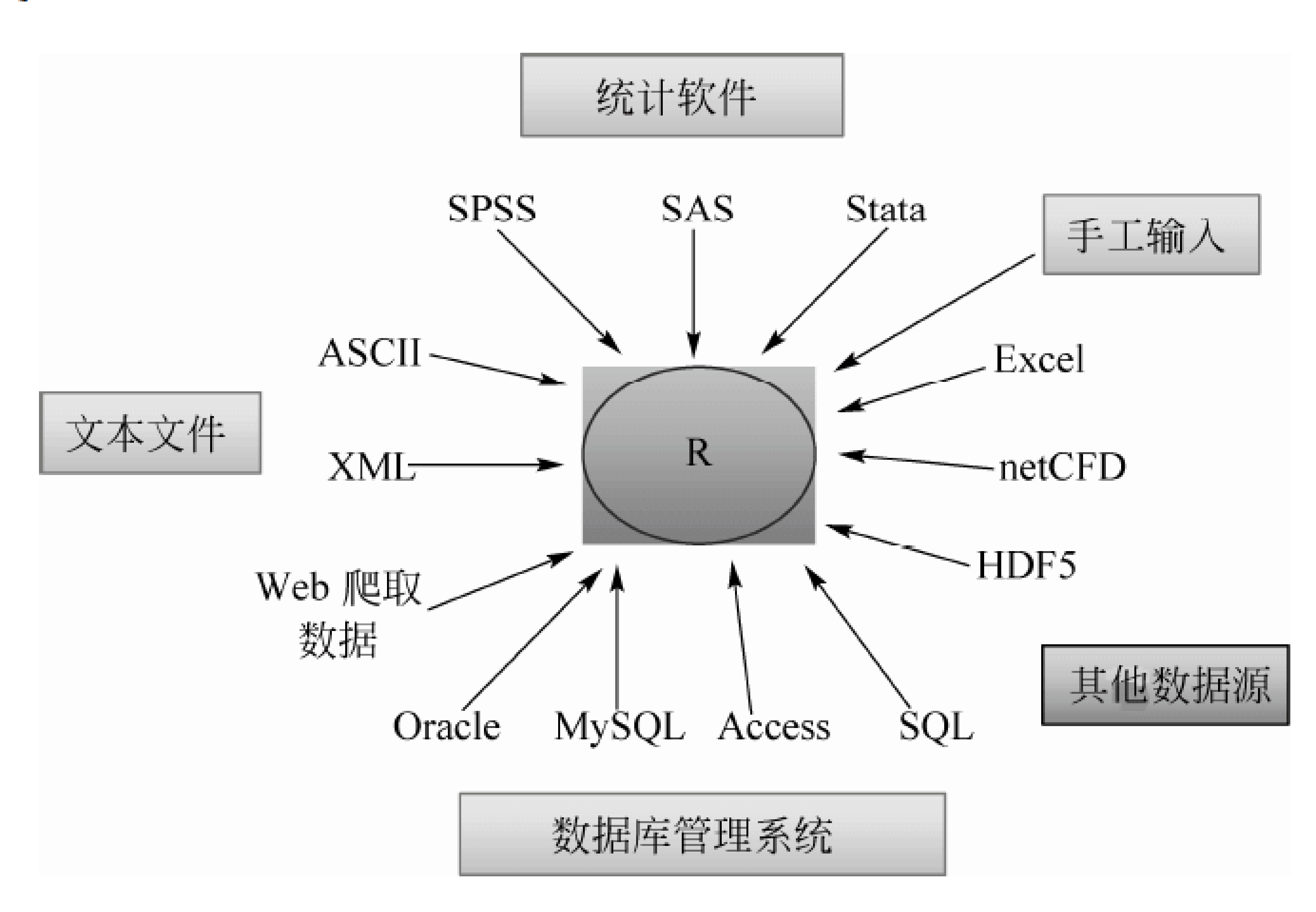


图 8-3-8　R 软件数据源

R 语言的数据结构包括向量（Vector）、矩阵（Matrix）、数组（Array）、数据框（Data Frame）和列表（List），通过这些数据结构来存储数据。

向量（Vector）为一维阵列。矩阵（Matrix）为二维阵列。数组（Array）与矩阵类似，区别是可以支持多维列。数据框（Data Frame）同样与矩阵类似，不同之处是不同列可以存储不同的数据类型，与数据库中的数据表类似。以上数据结构支持数字、字符和逻辑 3 种数据类型。

向量（Vector）、矩阵（Matrix）、数组（Array）、数据框（Data Frame）和列表（List）的数据结构如图 8-3-9 所示。

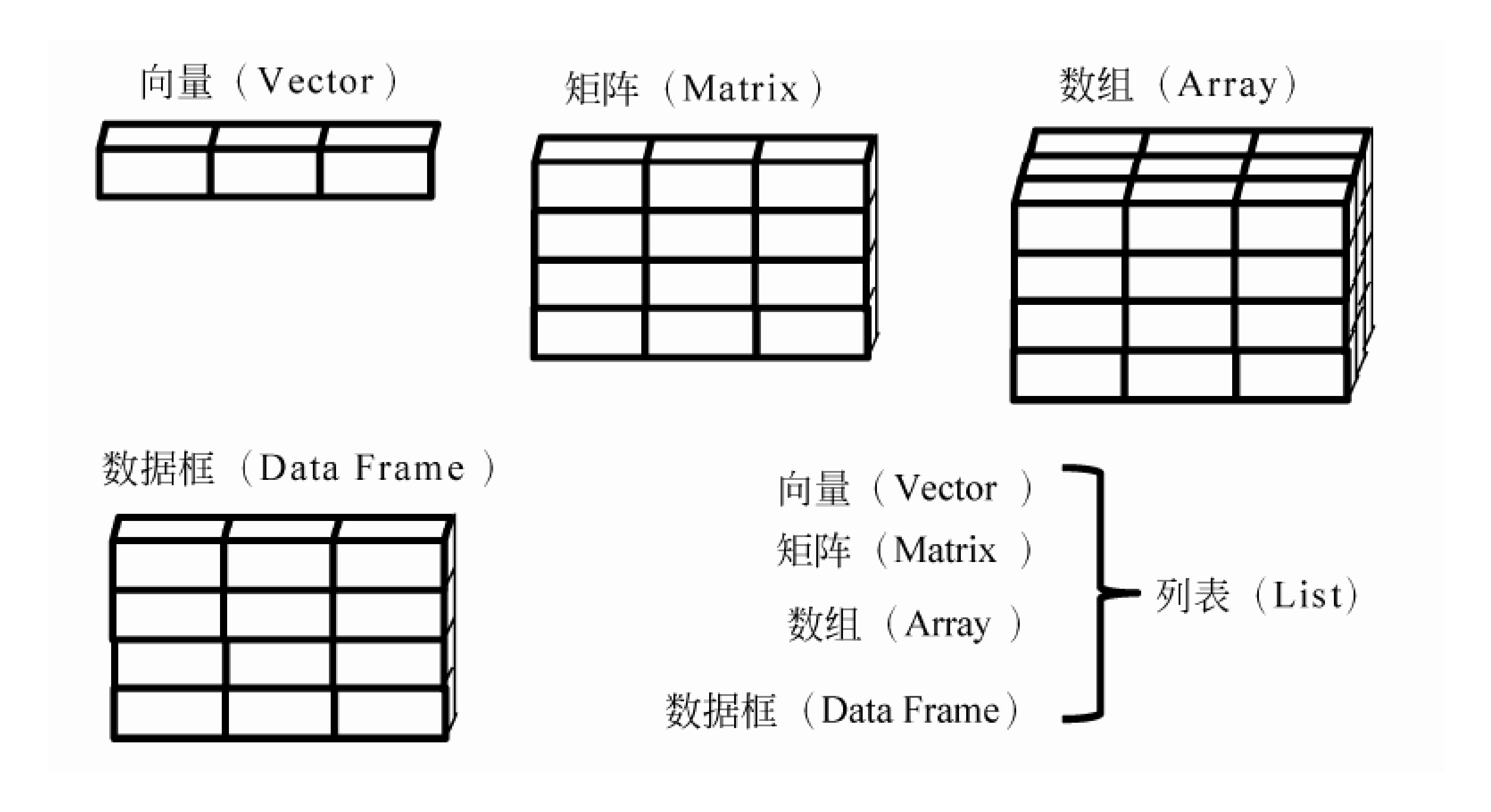


图 8-3-9　R 语言数据结构

R 软件的最大特点是支持回归分析。世界是普遍联系的，通过回归分析，可以发现事物之间的联系。

回归分析包括线性分析和多元分析两种类型。线性分析比较简单，用于分析某个自变量对于因变量的影响。多元分析则是分析多个自变量对于因变量的影响。

线性分析的例子包括广告费用与销售额的关系、预估价格与实际销售价格的关系、设备使用年限与设备费用的关系，等等。

多元分析的例子包括卡路里消耗与运动时长、平均速度（mph）、年龄、性别、平均心率、身体质量指数(BMI)的关系，人口自然增长率与国民总收入（亿元）、居民消费价格指数增长率（CPI）、人均 GDP（元）的关系，病虫与蛾量、卵量、降水量、雨日、幼虫密度的关系，等等。

统计分析工具能够帮助用户构建数据模型，集成不同来源不同格式的数据，让统计分

析结果更加直观，发现数据背后隐藏的规律，但在工具使用的过程中，往往需要分析人员结合自身经验来判断数据模型设计是否合理。

此外，R 软件等统计分析工具虽然具有强大的统计、预测分析以及数据可视化能力，但是其数据基础仅仅是处于大规模数据内部的样本数据，因此需要将统计分析软件与 Hadoop、HBase 等大规模数据管理软件结合起来。Hadoop、HBase 等大规模数据管理软件负责存取海量数据并形成统计分析软件所需的样本数据，R 软件等统计分析软件则专注于数据建模、统计以及数据可视化工作。

8.4 大数据展示技术

从多个维度、多个视角、全方位、直观地发现大数据背后隐藏的规律，相当于大数据挖掘的“最后一公里”。

“一图胜千言”，图形让人们更加直观地发现大数据背后隐藏的规律。随大数据展现技术可以从多个维度、多个视角、全方位、直观地发现大数据背后隐藏的规律，对于大数据服务具有非常重要的价值，相当于大数据挖掘的“最后一千米”。

按照数据展示的方式，将展示技术分为 Web 展示技术、GIS 展示技术以及移动客户端展示技术。

Web 展示技术采用 B/S 系统架构，B 就是 Web 浏览器（Browser），S 就是服务器（Server）。目前支持 B/S 结构的主流架构为 JEE 和.NET 两种，JEE 通过 JSP 进行动态网页的展示处理，而.NET 通过 ASP 进行动态界面的展示处理，JSP 和 ASP 返回给 Web 浏览器的结果都是 HTML。

GIS 是基于地理信息的展示技术，从空间角度展示，展示效果更加直观，更容易激发创造性思维。

移动客户端技术就是在移动终端上展示的技术，与桌面客户端相比，移动客户端的界面通常要小，因此与桌面客户端有不一样的展示要求。此外，移动智能终端还具有随身性，可以实时记录使用者的位置。定位技术与 GIS 技术结合，可以实时地展示人和物的运动轨迹。目前主流的移动客户端包括 Android（安卓）和 iOS 两种。

8.4.1 Web 展示技术

Web 展示的基础语言为超文本标记语言（Hyper Text Markup Language，HTML）。HTML 是一种有 Web 浏览器解释并展示的语言，经过 5 次重大的修改，W3C（万维网联盟）推出了旨在使 Web 开发更简单高效的 HTML5。

HTML5 具有跨平台、自适应、即时更新等优点，受到谷歌、苹果等公司的支持，成为事实上的网络标准。

Web 技术架构主要包括由 Sun 公司推出的基于 Java 的 Java 企业版（Java Enterprise Edition，JEE）和微软公司推出的.NET 两种类型。JEE 的特点是基于跨平台的 Java 构建，可以引入各种开源技术和框架，而.NET 则是基于微软的自有平台构建的，开放性差。近年来，JEE 因其开放性和跨平台性而得以广泛应用，下面主要分析 JEE 架构。

JEE 架构基于 Java 虚拟机构建，而 Java 虚拟机屏蔽了操作系统的异构性，因而基于 JEE 架构构建的应用具有在多种不同操作系统平台运行的特点，理论上讲，Java 程序可以实现“一次编译，到处运行”的效果，而实际上，由于不同操作系统平台之间存在一些差异，因此基于 Java 的软件代码在不同的操作系统平台上还需要做一些适应性修改。当然，这些修改的工作量要比重新编写软件小得多。

JEE 技术架构由多种语言和工具组成，主要包括 Java、HTML、JSP、JavaScript、CSS、JavaBean、EJB 以及开源框架 SSH（Struts、Spring、Hibernate）等。JEE 总体技术架构如图 8-4-1 所示。

从图 8-4-1 的 JEE 总体架构可以看出：HTML、JavaScript、CSS、JSP 属于 View 层，基于 Web 中间件实现 Web 业务逻辑的 Servlet 属于 Control 层，而 JavaBean、EJB 则属于 Model 层。开源框架 Structs 位于 Control 层。开源框架 Hibernate 位于 Model 层，衔接关系型数据库和 Java 对象。Spring 开源框架是一个容器，用于管理 bean 对象，不属于 MVC 设计模式的任何一层。

如果对 JEE 表现层进行细分，那么 HTML 负责 View 层的界面展现，JavaScript 负责 View 层的界面逻辑控制，CSS 负责 View 层的界面模型，HTML、JavaScript、CSS 又按照 MVC 模式进行了细分。

目前，主流的 Web 浏览器包括微软公司的 IE、苹果公司的 Safari、谷歌公司的 Chrome 以及开源社区 Mozilla 的 Firefox。

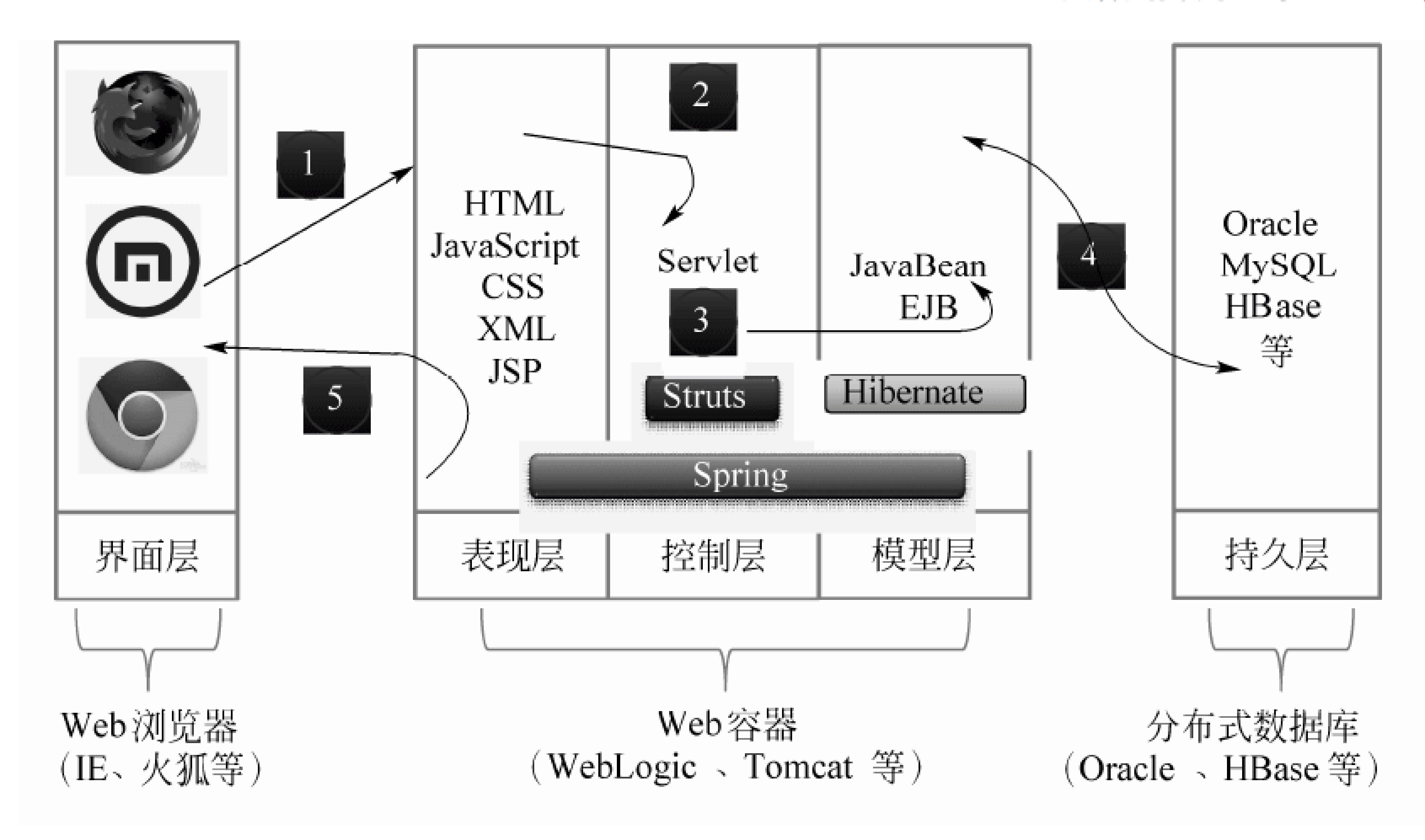


图 8-4-1　JEE Web 总体技术架构

Web 界面通常采用表格和图形两种方式来展示统计分析结果。表格形式展示的数据非常详细，但不直观、形象，图形方式则可以直观、形象地展现统计结果。FusionCharts，eCharts 等 Flash 控件，通过网页嵌入的方式，以 XML 格式的统计数据为输入，以图形方式展现统计结果。Flash 控件支持的图形包括 2D/3D 柱状图、曲线图、2D/3D 饼图、2D/3D 环图、区域图、堆栈图、联合图等。

下面是一个 Flash 控件显示统计结果的例子，简要说明了 JEE 架构中使用 Flash 控件展示统计结果的方法和过程。

首先，在 JSP/HTML 中通过 JavaScript 加载用于显示统计结果的控件（比如 FusionCharts），并在 JSP/HTML 界面上放置统计功能按钮，并通过预先定义的回调函数（CallBack）接收返回的结果。

其次，当浏览器页面发送统计请求后，请求会调用服务器端的 JavaBean 对象，JavaBean 对象再通过 JDBC 的方式从数据库（比如 MySQL）中获取统计数据，将统计结果拼装成 XML 文件并返回到预先定义的回调函数（CallBack）。

最后，将返回结果赋值给 Flash 控件后，执行界面刷新，就可以在 Web 界面上看到 Flash 控件展示的统计结果了。

以上就是一个借助 Flash 控件展示统计结果的简单过程，如果统计对象与空间信息有关，则可以借助 GIS 技术完成统计结果的展示。

8.4.2 GIS 展示技术

GIS 是 Geographic Information System 或 Geo-Information System 的缩写，中文为地理信息系统。

随着 GIS 技术的发展，GIS 在社会各行各业得到越来越广泛的应用。GIS 应用的领域可以分为社会公共事业管理、企业生产运营、人类生活 3 个方面。

GIS 在社会公共事业管理方面的应用主要是自然资源的管理，例如土地利用规划、森林管理、野生动物栖息地分析、河滨地带监测、自然灾害评估等。

GIS 在企业生产运营方面的应用主要包括营销渠道分析、通信网络分析、企业物流网络分析、市场分析等。

GIS 在人类生活方面的应用主要在交通出行方面，比如交通实时路况查看、旅游景点人流分析以及与 GPS 技术结合的自驾导航等。随着移动终端技术、移动通信技术以及互联网技术的不断发展，在人类生活方面将会产生越来越多的创新型 GIS 应用。

GIS 数据的特殊性在于其数据基础为地理空间数据。GIS 数据与他其数据管理过程类似，同样包括数据建模、数据维护、统计分析几个阶段。下面就分别介绍 GIS 在数据建模、数据维护、数据查询统计与展示 3 个方面涉及的关键技术原理和实现方法。

1. GIS 数据建模方法

GIS 数据建模方法包括矢量法和栅格法两种。矢量法擅长对离散数据建模，栅格法擅长对连续数据建模，在实际应用中需要根据建模对象的特点采用适用的建模方法。

1）矢量数据建模法

矢量法采用（*x*,*y*）坐标和点、线、面（多边形）代表空间要素，比如点可以代表某个污水井盖，线可以代表某条道路或者河流，面则可以代表某块菜地或者果园。矢量元素如图 8-4-2 所示。

矢量数据为离散型数据，可以不受分辨率的影响，不因为图形的放大、缩小或者旋转等而失真。

矢量数据分为拓扑和非拓扑两类。拓扑需要精确表达要素之间的空间关系，而非拓扑则不需要精确表达。因此，非拓扑数据比拓扑数据显示要快，并且可以用于不同的 GIS 软

件。ESRI 公司采用图层（Coverage）表达拓扑数据，采用 shapefile 表达非拓扑数据。

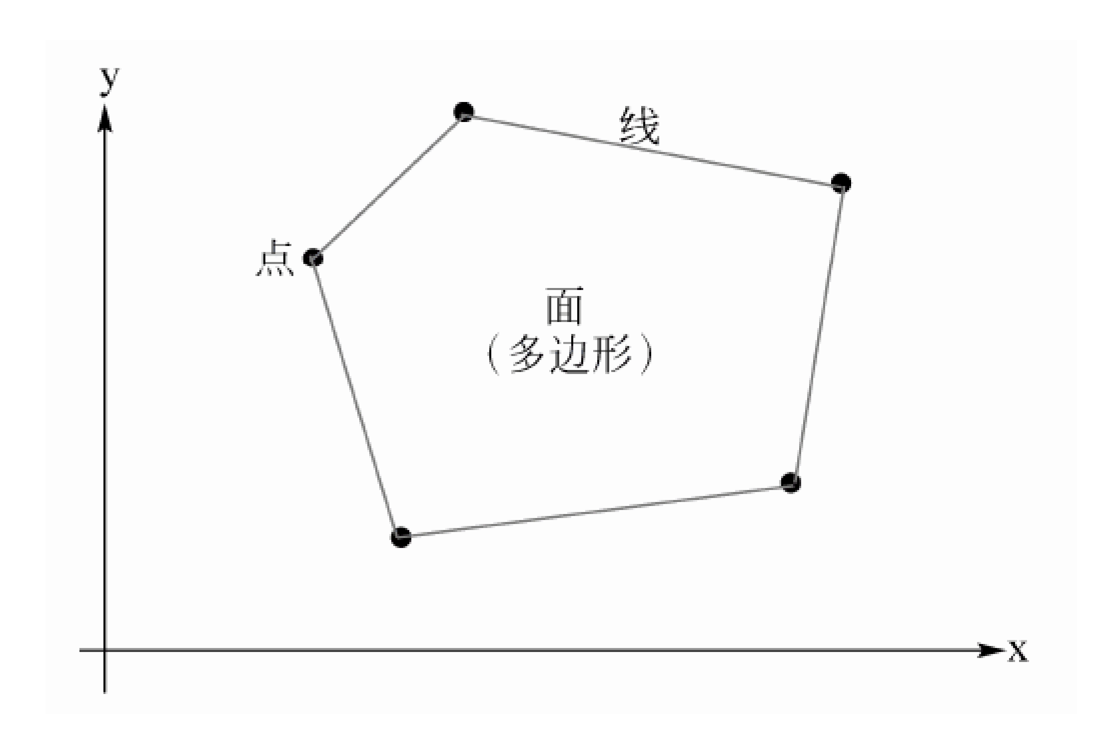


图 8-4-2　基于坐标系的矢量元素

2）栅格数据建模法

矢量数据模型虽然可以表达点、线、面（多边形）等离散型数据，但是对于海拔、降雨量、土壤侵蚀等连续性数据的表达并不理想。因此，人们提出了用栅格或者格网（Grid Cell）来表达连续性数据的方法。

栅格模型相关的概念包括高程、坡度、坡向等。高程（altitude），即高的程度，物体在给定的基准面（如地基、地面或海面）以上的垂直高度。栅格数据与矢量数据的对比关系如图 8-4-3 所示。

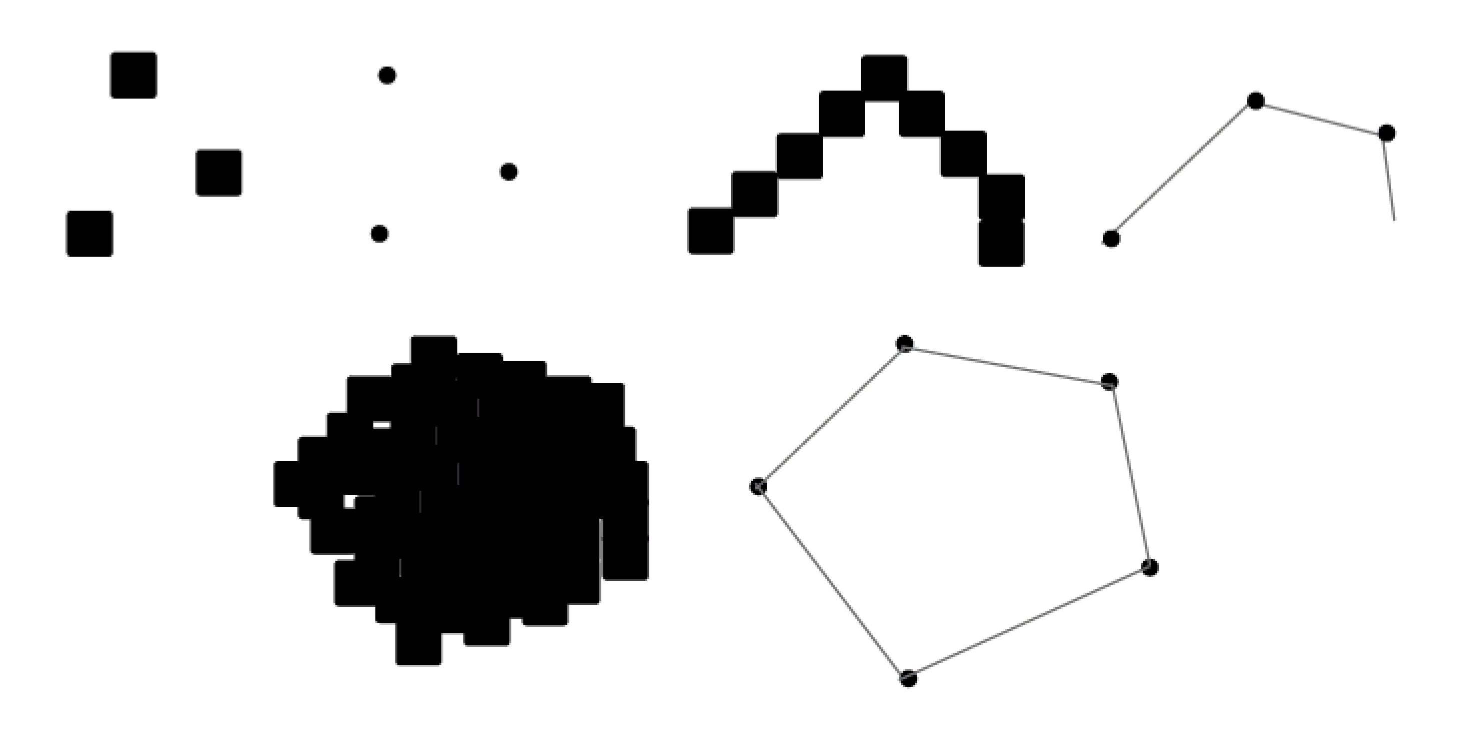


图 8-4-3　栅格数据与矢量数据的对比

从图 8-4-3 可以看出，栅格数据与矢量数据在表达点、线、面的方式上是不同的，栅格数据通过方块（格网元素）来表达各种图形形状，而矢量数据则通过不同点之间的连接关系来表达图形形状。栅格数据与矢量数据可以相互转化。

矢量数据以对象为描述基础，栅格数据则以域为描述基础。用栅格描述的对象有卫星影像、扫描地图、图形文件等。GIS 软件通常既支持矢量数据又支持栅格数据的显示。

栅格数据模型也称为格网。格网由行、列、格网单元组成，行列从左上角开始。格网单元的大小决定了栅格数据模型的分辨率。单元依序编码是栅格编码的一种方式，如图 8-4-4 所示。

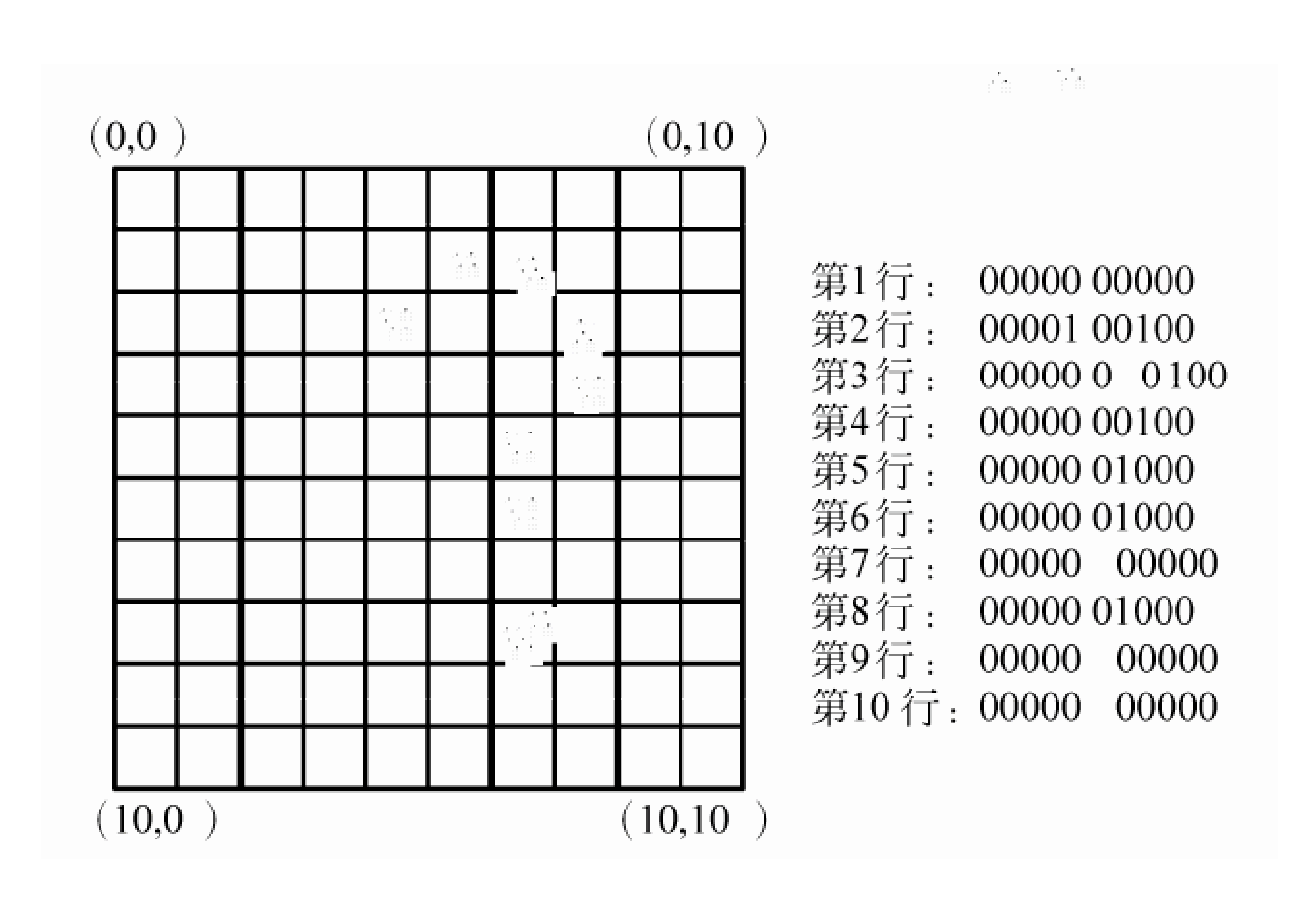


图 8-4-4　栅格的单元依序编码方式

从图 8-4-4 可以看出，栅格是以左上角为坐标原点的。在图 8-4-4 的右侧，栅格通过 1 或者 0 来表示网格单元内部是否有数据。在上面的图形示例中，通过栅格表达了一个像？的图形。

网格单元的大小决定了栅格数据模型的分辨率，栅格越小，栅格数量越多，表达的图像分辨率越高，当然占用的存储空间也就越大。

在网格中存储多个属性值的方法是采用单元 ID 作为网格值，这样就可以通过该单元 ID 获取到多个属性值。栅格数据模型难以表达空间要素的精确位置时，要用矢量数据模型来表达。

3）空间数据与属性数据

地理数据分为空间数据和属性数据两种类型，空间数据与地图要素的几何特征有关（比如坐标系、x 值、y 值等），而属性数据则描述空间要素的特征，比如某个空间区域的人口数量、不同种族、性别、教育程度的人口数量细分等。

地理数据的原点是本初子午线和赤道的交点，经度相当于坐标系统的 x 值，以本初子

午线为中界线，向东或者向西 0° ～180° ，纬度相当于坐标系统的 y 值，以赤道为中界线，向南或者向北 0° ～90° 。

地理空间数据模型将空间数据和属性数据分开存储。空间数据采用图形文件存储，用文件管理系统来管理，Windows 操作系统通过文件分区表（File Allocation Table，FAT）进行文件的管理，而 Linux 操作系统则通过 Ext3（延伸系统）等管理文件。属性数据则采用 Oracle、MySQL 等关系型数据库来存储和管理。

2. GIS 数据维护方法

空间数据的维护方法是采用人工和自动（比如卫星数据）相结合的方式，原因是空间数据是分散的，如果采用人工方式，一来是数据的准确性难以保证，二来是数据的采集成本高。移动互联网时代，空间数据的采集可以采用众包/众筹模式，发动全社会力量，从而解决了因空间数据分散而难以采集的问题。

属性数据的维护包括数据的输入和校验。数据输入可以采用人工方式，也可以采用外部文件导入的方式。属性数据的校验分为两种，一种是唯一性和非空校验，另一种是数据准确性校验。属性数据可以以地图单元符号为关键字查询关系型数据库获得。

3. GIS 数据查询统计与展示

地图是刻画空间的一种方式。地图包括图名、图例、指北针、比例尺、文字说明、图廓、空间要素等。比例尺、准确度、精确度是衡量地图质量和能力的几个重要指标。

地图可以通过图层表示不同层次的空间元素。不同层次的地图比例尺是不一样的，地图上比例尺通常以厘米为单位，比如世界地图中的国、省、市、街的比例尺可以分别为 1:1000km，1:50km，1:5km，1:200m，就是指地图上的 1cm 分别代表 1000km、50km、5km 和 200m。

地图可以分为普通地图和专题地图。普通地图包括边界线、水文、交通、等高线、居民点、土地覆被等；专题地图包括人口密度分布、网络流量分布、实体渠道分布等。主流的地图软件包括谷歌地图、百度地图等。ArcGIS 是开发地图软件的一款功能强大的工具软件。

可以通过不同的颜色或者颜色的深浅来表示不同区域的特征，例如分地域统计上网流

量，直观展示不同地域在某时间段的总体流量情况，如图 8-4-5 所示。

空间数据查询是通过地图要素的操作从地图上检索数据的过程。可以使用指针、图形或者地图要素之间的空间关系来选择地图要素。用于查询的空间关系包括包含、相交、邻近。通常采用空间数据和属性数据结合的方式进行查询，比如查询建筑物周围半径为 1km 的树木数量，或者查询某个基站周围半径为 3km 的行政村个数。

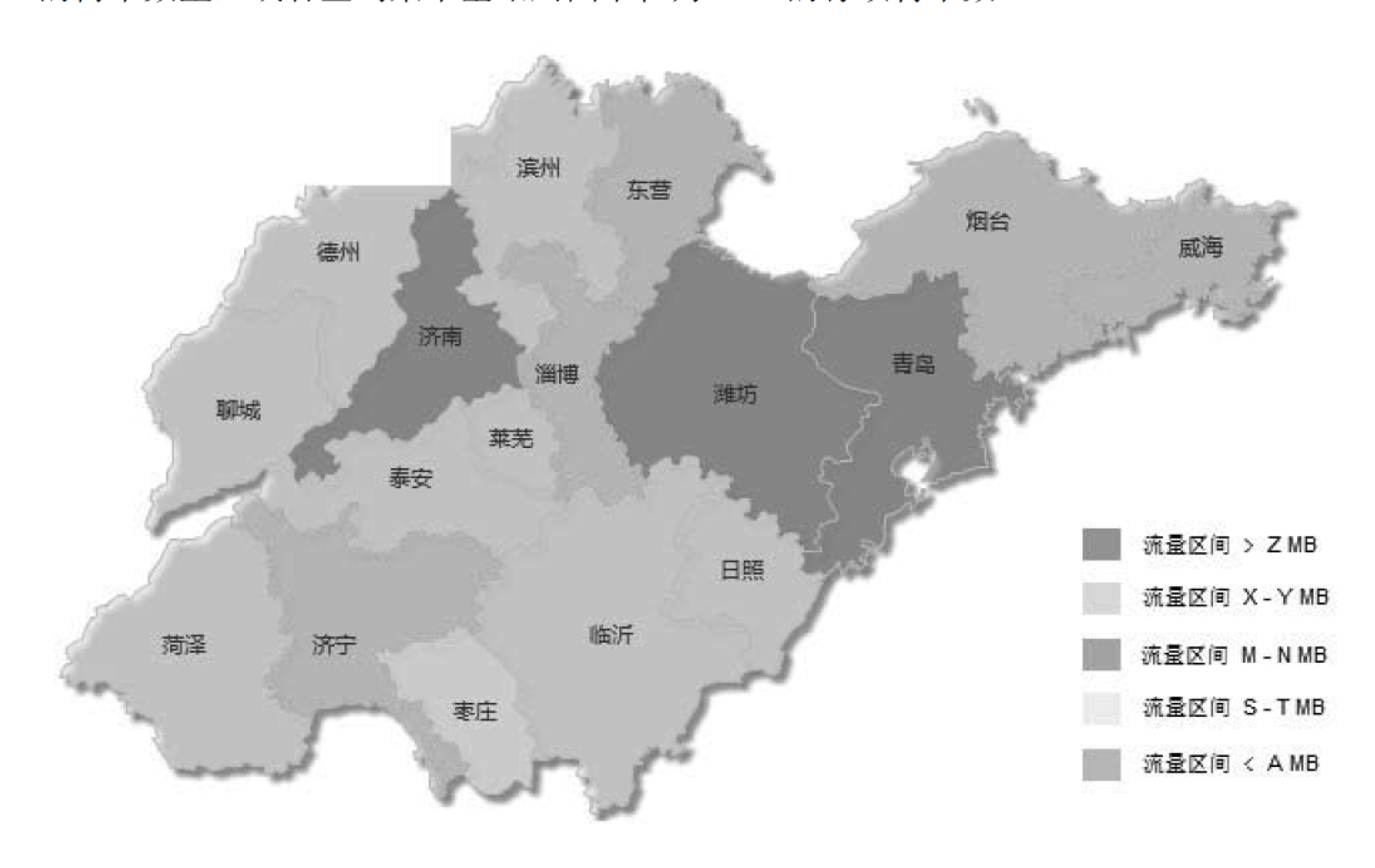


图 8-4-5　分省统计上网流量

随着移动通信技术（3G/4G）、全球定位系统（Global Position System，GPS，借助通信卫星定位）等技术的发展，基于位置的服务（Location Based Service，LBS）也得到快速的发展。

LBS 是定位技术与 GIS 结合的结果。通过定位技术与 GIS 的结合，可以开发出丰富的 LBS，比如企业可以通过 LBS 来推送商品促销信息，用户可以通过 LBS 来找到附近的商家、加油站、银行网点等，公安部门则可以借助 LBS 找到嫌疑犯的行踪。人们在自驾旅行时用的汽车导航系统是一个典型的例子，导航系统可以通过 GPS 来实时定位汽车的位置，并根据导航目标进行行驶指示，同时，导航系统借助 3G/4G 等移动通信网络下载最新的地图数

据，当导航系统中没有目标位置时，也可以去服务器上检索。当下流行的导航软件包括高德导航、腾讯导航、百度导航等。

8.4.3 移动客户端展示技术

移动终端是指可以移动的终端设备，包括手机、平板电脑、导航仪等，移动客户端则是基于移动终端开发的、C/S 结构的软件。

移动终端的两大主流操作系统是 Google 开源的安卓（Android）系统和苹果公司的 iOS 系统。据数据分析公司 Strategy Analytics 统计，2014 年第三季度，安卓市场占有率为 83.6%，而 iOS 市场占有率为 12.3%。尽管 iOS 的市场占有率不及安卓，但是 iOS 的用户往往是一些高价值用户，因此 iOS 用户的总体收入并不低。

与桌面终端不同的是，移动终端包括移动通信模块和移动定位功能，这使得人们可以摆脱时间和空间限制，实现 5 个 A（任何时间、任何地点、任何终端、任何网络、任何数据）的自由联通。此外，移动终端的定位功能使得无线电台（基站、卫星等）和移动终端之间保持实时的位置更新，移动终端可以结合 GIS 系统构建各种各样的 LBS 应用，比如汽车导航、位置营销等。

由于移动终端的随身性，在大数据时代，基于移动互联网的大数据服务将会变得越来越重要，企业可以借助移动互联网提升营销能力和服务能力。为了说明移动客户端的实现方式，下面重点分析安卓平台的架构。

安卓操作系统由谷歌公司领导和开发，主导思想是开源和开放，这与谷歌在搜索等领域采用的开源和开放思想是相似的，只有通过开源和开放，才能调度全社会的开发力量，为用户提供各种创新型应用，才能让安卓技术具有源源不断的发展动力和持久的生命力。

安卓系统是从 Linux 平台发展而来的，由操作系统、中间件、用户界面和应用软件组成。安卓系统架构如图 8-4-6 所示。

从图 8-4-6 可以看出，安卓系统自上而下分为 5 个部分，分别为应用、应用框架、库、安卓运行时以及 Linux 内核。

1. 安卓系统应用

包括电子邮件、短信、地图、浏览器、通讯录等。

2. 安卓系统应用框架

安卓系统应用框架包括 Activity Manager、Content Providers、Resource Manager、Location Manager、Notification Manager 等。



图 8-4-6　安卓系统总体架构

（1）Activity Manager（活动管理器）：控制应用的生命周期，在用户导航的时候，当用户使用其他应用时维护一个回退栈，即将暂不使用的应用放到栈中，这样再使用栈中的应用时，无须重启该应用，提高了用户访问应用的效率。

（2）Content Providers（内容提供商）：这些对象封装了需要在应用之间共享的数据，比如通讯录。

（3）Resource Manager（资源管理器）：资源是程序中那些无须编码的东西。

（4）Location Manager（位置管理器）：安卓系统总是能够掌握自己所在的位置。

（5）Notification Manager（通知管理器）：比如消息、任务、告警等事件，以最优雅的方式展示给用户。

3. 安卓系统库

安卓系统库支持上层应用，主要包括 Surface Manager、Graphics、Media、SQLite、Webkit、Libc、FreeType 等。

（1）Surface Manager（外观管理器）：使用与 Vista 类似的组合窗口管理器技术，并且更加简单。不是直接基于屏幕缓冲区绘制，而是将绘图命令深入到后台的位图，然后与其他位图合并后显示给用户，这使得系统可以构建任何有趣的显示效果，比如透视窗体与动画效果。

（2）Graphics（图形）：二维和三维的元素可以合并为一个用户界面。该库会使用具有 3D 功能的硬件。SGL 为底层的 2D 图形引擎。

（3）Media（媒体）：可以支持多种格式的语音和视频媒体，比如 AAC、AVC（H.264）、H.263、MP3、MPEG-4 等。

（4）SQLite（SQL 数据库）：支持轻量级的数据库引擎 SQLite，该数据库同样在 Firefox 和苹果 iOS 中使用，可以实现应用数据的持久化存储。

（5）WebKit（浏览器引擎）：用于显示 HTML 内容，同样的引擎也应用于谷歌的 Chrome 浏览器、苹果的 Safari 浏览器、诺基亚的 S60 平台。

（6）Libc（C 系统库）：针对标准 C 系统库的派生实现，针对嵌入式 Linux 进行了调整。

（7）FreeType：针对位图和矢量字体的渲染。

4. 安卓系统运行时

Dalvik 是为移动终端专门打造的虚拟机。Dalvik 不同于 Java 虚拟机，虽然 Android 应用采用 Java 语言，但是当 Java 应用编译成.class 文件后，还是需要通过工具将其转换成能够在 Dalvik 上运行的 dex 文件。

安卓 5.0.1 版本的 SDK 对于网络、多媒体、图形、数据库等的支持如图 8-4-7 所示。

安卓 5.0.1 版本的 SDK 对于显示、通信网络、系统、安全等方面的支持如图 8-4-8 所示。

当前，安卓应用主流的开发环境为 Eclipse+ADT+SDK，Eclipse 是集成开发环境，安卓开发工具（Android Development Tools，ADT）是基于 Eclipse 上实现快速开发安卓应用的插件，开发测试的工具包是安卓 SDK。由于 Eclipse 是一个通用的应用开发平台，在支持安卓应用开发方面存在一些不足，为此 Google 推出了自己的开发工具 Android Studio，

与基于 Eclipse 的安卓开发工具相比，Google 的 Android Studio 易用性更好。

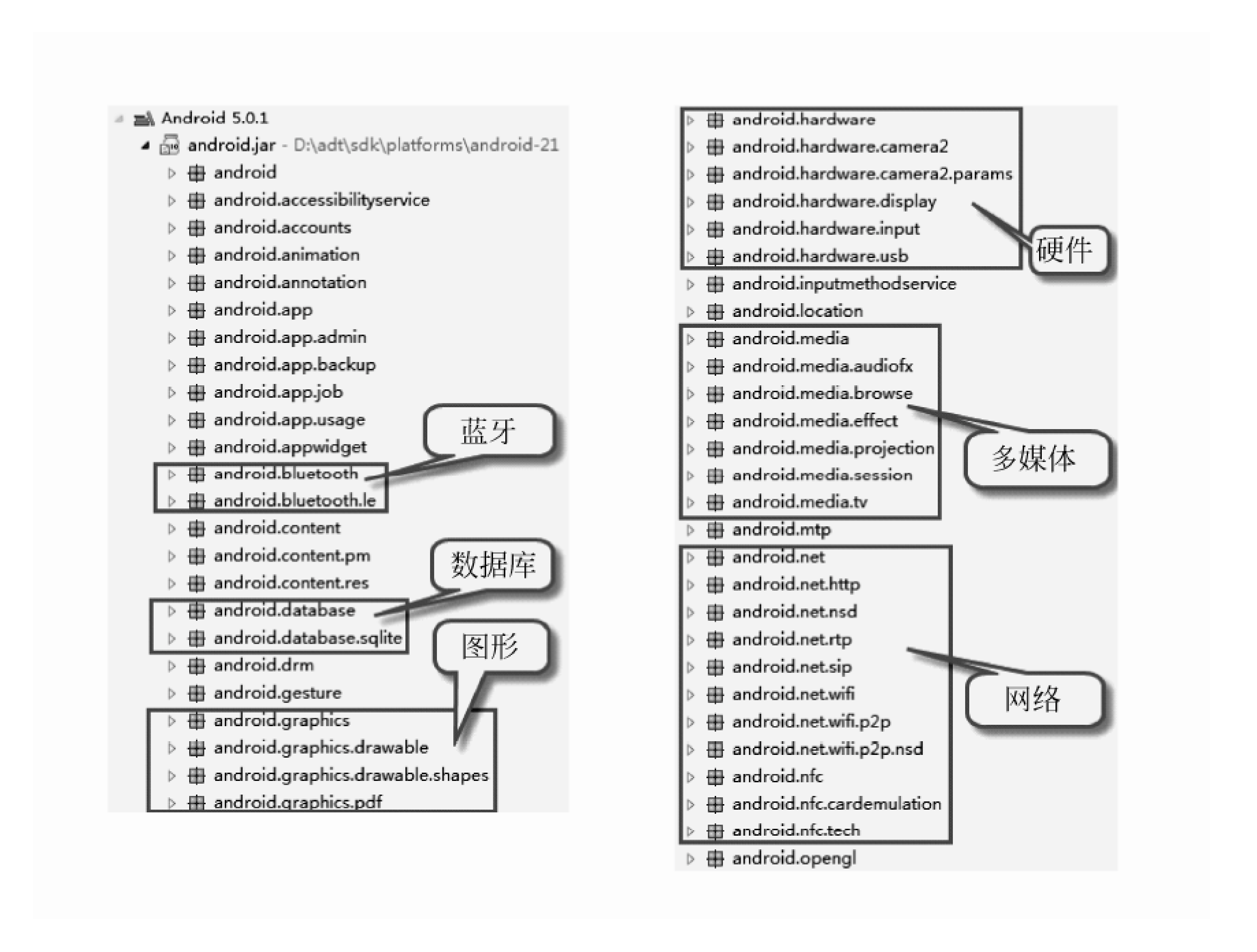


图 8-4-7　Android SDK 对于网络、图形、数据库、多媒体等方面的支持

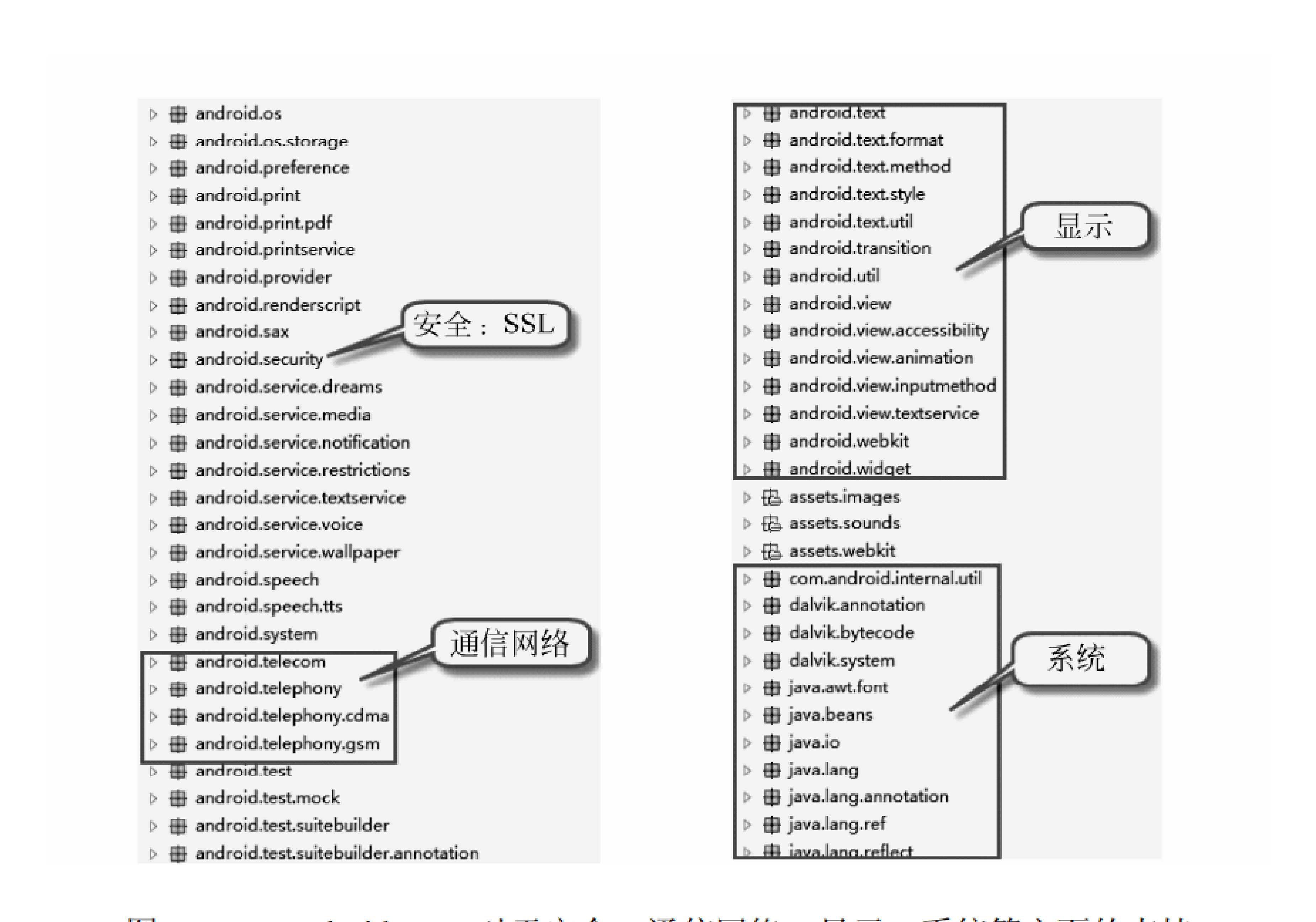


图 8-4-8　Android SDK 对于安全、通信网络、显示、系统等方面的支持

苹果的 iOS 系统采用了与谷歌的安卓系统不同的商业模式，安卓系统开源开放，强调开放和共同参与，而苹果 iOS 系统则正好相反，iOS 是一款全封闭的移动终端操作系统，iOS 操作系统完全由苹果公司自己控制。苹果公司之所以这样做，完全是为了通过对于操作系统细节的管控，实现完美的用户体验。

8.5　主要内容回顾

企业要实现基于大数据的运营，大数据相关技术起到至关重要的作用。

不同于支持事务型应用的技术，大数据技术具有自身的特殊性，分别体现在数据存储、数据分析以及数据展示 3 个方面。

1．大数据存储技术总结

在数据存储方面，大数据具有数据容量大并且随着时间的变化数据量持续增长的特点，因此要求数据库存储系统能够做到存储空间的线性扩展，这样就不会出现因为存储空间不足而影响系统整体性能的问题。

以 HDFS 为代表的分布式文件系统，将数据文件放入主机内部的磁盘上，而不是像传统方式那样放到独立的磁盘阵列中，同时以数据块为存取单位而不是字节，从而提高了数据存取的效率。每个数据存储节点都是分布式存储系统的一个节点，节点之间互不影响，因此可以做到存储空间的线性扩展。

当然 HDFS 也存在着不足，由于各个存储节点都是独立的，如果想关联多个存储文件的内容，需要遍历多个主机节点，效率很低，因此主要适合面向数据表之间关联度低的分析型应用。

基于关系代数理论的关系型数据库存储系统与分布式文件系统不同，关系型数据库存储系统基于单机模式发展起来，擅长多个数据表之间的关联分析，但关系型数据库的不足之处是数据库扩展性差，虽然采用集群方式提高了数据库系统的扩展性，但是受限于集中控制模式的限制，数据库扩展性非常有限，并且随着集群规模的不断增大，存储系统整理性能急剧下降，无法做到线性扩展。

2. 大数据分析技术总结

在数据分析方面，从数据分析的实时性角度分为离线分析和实时流式分析两类，从数据分析的模式角度分为大数据 MapReduce 分析和关系型数据库分析两种类型，从数据建模角度分为大规模数据分析和基于样本数据的分析两类。

离线分析技术适用于实时性要求不高的场景，特点是支持的数据规模大。实时流式分析可以快速地完成数据的统计，但是仅仅适合于完成海量数据某一个侧面的计算，比如用户偏好画像、搜索关键字统计等。

大数据计算模型以 MapReduce 最为经典。MapReduce 算法由谷歌公司发明，好比高等数学里面的微积分。首先将大文件“微分”为多个小的数据块并存入 HDFS 集群中，然后再通过 MapReduce 完成对“微分”数据的“积分”。Map 负责以映射的方式提取分散在大数据集群中的数据项，Reduce 则负责对排序后的统计数据进行聚合（求和、求均值等）输出。MapReduce 计算模型特别适合对分布式文件系统中的统计分析。

MapReduce 计算模型之所以能够满足海量数据的统计，根源在于被统计文件虽然规模大，但是采用列式存储方式，原始数据具有共同的数据特征，而关系型数据是按行存取的，每一行中不同列的数据特征都不一样，要完成数据的统计需要扫描所有行，因此面向海量数据时的统计效率低，只能通过分区、索引等方式将数据规律性布放，提高数据的存取效率。

尽管 MapReduce 计算模型非常强大，但是如何实现统计功能需要编程实现，而开源工具 R 软件采用命令行方式，可以快速完成数据建模、统计分析以及可视化工作。R 软件的优势是能够快速调整模型、快速见到分析结果，不足之处是对于海量数据的分析能力差，因此需要将 Hadoop/MapReduce 计算模型与 R 软件结合起来，R 软件侧重基于样本数据构建分析模型，而 MapReduce 则侧重于为 R 软件提供样本数据。

3. 大数据展示技术总结

“一图胜千言”，大数据分析结果如果以表格形式来展现统计报表结果，则很难发现数据背后隐藏的规律，原因是客户世界中不同事物之间的联系并不是线性的，而是网状的。

按照分析结果展示的形式，将大数据展示技术分为 Web 展示技术、GIS 展示技术以及

移动客户端展示技术 3 种类型。

Web 展示技术是最简单、最传统的展示技术，它借助 Web 浏览器和图形控件展示统计分析结果。Flash 控件是当下流行的 Web 展示技术，Flash 控件提供了统计结果数据的注入接口，可以以 2D、3D 形式展示柱状图、饼图、地图等多种统计图形。

GIS 展示技术侧重于展示具有空间特性的对象，比如河流湖泊、交通路线、通信网络等。基于 GIS 技术展示统计结果，更具有现实感，更能够激发创造性。GIS 展示技术与定位技术相结合，会形成多种基于位置的创新型应用。

移动客户端展示技术主要考虑移动终端屏幕大小和用户位置信息，由于移动终端的随身性，与 GIS 相结合则能够直观地掌握用户的运动轨迹，可以定位地理空间的目的地以及到达该目的地的路线。

附录 A
重点概念及其定义

1. 服务型企业、制造型企业

服务型企业以提供无形的服务为主，制造型企业则以提供有形的产品为主。例如，电信运营商提供信息通信服务、商业银行提供存贷款服务，它们都属于服务型企业，而像农业机械设备制造厂、飞机制造厂等企业，为客户提供的是有形的零件或者设备，因此属于制造型企业。

本书中所述的理论和方法主要适用于服务型企业。

2. 架构、框架

架构侧重某个系统的全局和整体，比如盖房子需要在架构设计方案中明确骨架及其连接关系。框架则是架构中某个特定的部分，比如一扇窗户是一间房子的一部分，那么窗户的造型设计可以认为是一个框架。

3. 业务、服务

广义的业务是指企业生产经营中需要处理的事务，狭义的业务是指企业能够为客户提供的能力，例如电信运营商为客户提供的业务包括语音业务、数据业务等。广义的服务是为对方做事，并使对方从中获益的一种有偿或者无偿的活动，狭义的服务是指一种抽象的能力，由于服务是抽象的，可以摆脱资源不灵活的限制。例如，电信运营商把各种资源抽象成服务，这些服务再通过封装，形成各种各样的能力。服务的目标不同，其关注点就不同，可以将服务分为面向客户的服务和面向资源的服务，面向客户的服务主要关注市场价值属性，而面向资源的服务则主要关注使用属性。

4．职能、过程

职能对应英文的 Function，也可以理解为功能，过程对应英文的 Process，由于历史原因，Process 在许多情况下被翻译成“流程”并被广泛使用。实际上，将 Process 翻译成“过程”更为合适，“流程”的英文单词应当为 flow。过程是从时间、顺序上看待事物，相当于思维时空在时间轴上的投影。流程则是更一般的用语，比过程更具体。为了统一术语，本文将 Process 统一翻译成“过程”，例如 Business Process Framework 翻译成“业务过程框架”而不是“业务流程框架”。

5．业务过程、应用、功能

业务过程是采用面向过程的思维方式对业务活动的定义。业务过程从动态角度定义业务活动，在业务活动执行过程中会形成静态信息，信息是业务的“概念”，对信息需求进行建模会形成概念模型。

广义的应用是用户使用信息系统获取到的服务，比如，腾讯公司推出的微信就是一种移动即时消息应用，许多软件服务可以放到应用商店中，供用户下载、安装和使用。狭义的应用为介于业务和技术之间的 IT 能力，应用的主要目的是为业务人员和技术人员提供中介物，业务人员可以提出开发一个支持促销活动的应用，而技术人员则需要根据应用需求，设计、开发并实现支持促销活动的应用。

与应用相比，功能更加具体和细微，比如微信应用中包括多个具体的功能点，如微信内容浏览功能、小视频上传功能等。

6．业务过程框架、信息框架、应用框架、集成框架

业务过程框架是对企业所有企业活动的管理。时间维度包括战略、基础设施、产品、运营支持与就绪、服务履行、服务保障、服务计费和收入管理几个阶段。空间维度包括市场、客户、产品、服务、资源、供应商、合作伙伴几个部分。

信息框架是对企业信息进行管理。信息框架同样可以分为市场、客户、产品、服务、资源、供应商、合作伙伴几个域，然后再使用分层的方式进行细分。信息框架对应概念模型，概念模型就是在需求分析和设计时使用的实体关系模型，实体关系模型是从静态结构的角度对业务需求的刻画。

应用框架是对应用管理的一种模式。应用框架与业务过程框架非常相似，是业务过程框架和信息框架向 IT 实现的一种收敛。应用框架体现了软件复用的思想，将一些通用的应用提取出来，比如日志管理、安全管理等。此外，应用框架还包括 IT 基础设施通用能力部分，比如业务流程管理、企业服务总线、工作流等。

集成框架是对业务过程框架、信息框架、应用框架的集成。业务过程框架提供动态的业务服务，信息框架提供静态的实体服务，应用框架则提供公共的通用服务。在业务层面，集成框架通过业务服务（Business Service）的方式实现业务过程的标准化和集成，支持面向价值网络的企业。在技术层面，集成框架则体现了系统或者平台之间的集成接口，包括 API、数据库、Web Service、文件等集成接口方式。

7. 面向操作的事务型应用、面向决策的分析型应用

从应用对数据的处理行为角度看，操作型应用对数据主要执行增加、删除、修改动作，而分析型应用对数据主要执行查询、统计动作。从应用支撑的目标角度看，操作型应用主要满足企业日常的建设、生产、运营和管理需要，比如企业的客户关系管理系统、计费账务系统、办公自动化系统等，分析型应用主要满足企业生产经营过程中战略、战术、操作层面的决策。从数据的特点看，操作型应用通常是操作单个数据，数据总量小，而分析型应用则通常操作批量数据，数据通常是长期积累的历史数据，数据总量大。

8. ODS、数据仓库、数据集市、商业智能、大数据

ODS（Operational Data Store）是操作型数据存储，数据来源于面向操作的事务型应用，存储内容为事务细节数据，组织通过对来自不同业务系统数据的采集与整合，实现数据查询和统计分析功能。ODS 与操作型应用产生的数据分开存储和管理，可以降低业务系统的查询压力。

数据仓库（Data Warehouse）是对大量数据存储的一个形象化概念，意味着对多种数据的存放。数据仓库之父比尔•恩门（Bill Inmon）将数据仓库定义为：在企业管理和决策中面向主题的、集成的、与时间相关的、不可修改的数据集合。可见，数据仓库中的数据是面向不同目标主题的历史数据的集合。

数据集市（Data Mart）是面向主题的数据集合体，就好比现实生活中的集市，按照不同的服务功能分为蔬菜、服装、肉类、生活用品等不同的专区，人们可以根据自身需求去不同的专区购买商品。数据集市是在数据仓库的基础上，按照特定的数据分析需求，对数

据进一步的汇总形成的，数据集市可以让数据更加聚焦在某一主题，比如客户、产品、渠道、资源、供应商、人力资源、财务、资产等。组织的不同部门对于数据的关注点不同，数据集市可以更好地满足不同部门的特定需求。

数据挖掘（Data Mining）通过从大量的、不完全的、有噪声的、模糊的数据中提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识。典型案例就是啤酒与尿布的故事，通过数据挖掘可以发现平时不能察觉的客户购买行为。数据挖掘需要以大量历史数据作为分析基础，还需要构建数据挖掘所需的模型，数据挖掘方法包括回归分析、聚类、关联规则、特征、变化和偏差分析、Web 页挖掘等。

商业智能（Business Intelligence）更多的是一个概念，其目标是将数据转换为知识，帮助企业做出明智的生产与经营决策。商业智能需要利用数据仓库、数据挖掘、在线分析处理等技术来实现。

大数据（Big Data）的特征是数据规模大、类型多样、产生速度快、面向价值。数据规模大是一个形容词，业界也称之为海量数据。类型多样意味着不仅仅包括结构化数据，还包括文本、图片、语音、视频、邮件等半结构化和非结构化数据。产生速度快是随着移动互联网和物联网的飞速发展，个人以及各种传感器成为数据产生的重要源头，数据产生的速度更快了。面向价值是指大数据的目标是发现价值。

9. OLTP、OLAP

在线事务处理（On-Line Transaction Processing，OLTP）包括增加、删除、修改等数据维护功能。OLTP 的特点是事务性，事务的特点是 ACID，即原子性（Atomicity）、一致性（Consistency）、隔离性（Isolation）、持久性（Durability）。事务要求保证操作的完整性，操作要么成功，要么失败，没有中间状态。

在线分析处理（On-Line Analytical Processing，OLAP）强调数据分析结果反馈的效率，许多时候用户希望能够快速地看到对大量数据统计的结果，“在线”体现了数据获取速度的要求，比如电信运营商的无线网络建设者需要快速掌握基站的数量、分布、数据流量区间等。OLAP 应用通常采用分区、集群等数据库技术来提高数据分析的效率，通过系统后台创建中间表的方式让用户快速看到数据统计的结果。典型的 OLAP 应用包括关于销售、市场、管理报表、BPM、预算与预测的商务报告。基本的多维分析操作有钻取（Drill-up 和 Drill-down）、切片（Slice）和切块（Dice）以及旋转（Pivot）等。

10．实时、准实时、非实时

这三种叫法都是从系统的处理和响应时间角度出发的。实时任务的响应时间通常是在1s 以内，对于用户来说几乎没有感到时间的延迟，因此称之为“实时”，其实绝对实时是不存在的；准实时任务的响应时间比实时的要长，通常在几分钟之内；非实时是与实时相对的叫法。

11．结构化数据、半结构化数据、非结构化数据

结构化数据是具有共同特征（比如数据类型、长度等）的数据集合，例如，以二维表格形式存储的个人信息（姓名、年龄、身高、体重等）属于结构化数据。

非结构化数据无法以二维表格形式管理，通常以文档、图片、录音、视频等形式存在。

半结构数据则介于结构化和非结构化数据之间，其主要特征就是数据定义（又称为元数据）和数据内容是合在一起的，比如 Web 网页、邮件等。目前结构化数据主要以关系型数据库存储和管理，更容易统计和分析，而非结构化数据则通常需要先完成结构化工作。

12．交互存储区、集成存储区、近线存储区、归档存储区

以上存储区主要是从数据生命周期角度划分的，不同的数据满足不同的应用需求和管理要求。

交互存储区（Interactive Sector）的数据产生速度通常在几秒钟之内，数据的活性最强；集成存储区（Integrated Sector）通常存储 1 天、1 个月、1 季度或者 1 年的数据，集成存储区的数据来自于交互存储区，通常用于 OLAP 应用；近线存储区（Near-Line Sector）通常存储 3～5 年的数据，数据可以直接来自于交互存储区，也可以来自于集成存储区，通常用于数据挖掘应用；归档存储区（Archival Sector）的数据可能来自于集成存储区和近线存储区，数据存储年限为 5～10 年，通常是由于政策法规的要求而存储的，数据的查询频率很低。

参考文献

[1] 李福东.Frameworx 框架体系浅析.邮电设计技术，2014-02.

[2] 李福东.面向移动互联网的需求管理与应用架构模式研究.邮电设计技术，2014-11.

[3] 李福东.电信运营商业务财务一体化需求分析思路与方法研究.邮电设计技术，2012.

[4] 李福东,张何林,吴盛博.ITILv3 框架体系浅析，北京通信协会优秀论文集，2013.7.

[5] 于海澜.企业架构:价值网络时代企业成功的运营模式.北京：东方出版社，2009.6.

[6] CIO 时代网: http://www.ciotimes.com/ea/eatools/ 200911301124.html。

[7] Scott A Bernard.An Introduction to Enterprise Architecture Third Edition.Bloomington：Author House，2012.

[8] 陈威如,余卓轩.平台战略——正在席卷全球的商业模式革命.北京：中信出版社，2013.1.

[9] 腾讯开放平台，2013 中国互联网开放平台白皮书.腾讯科技（深圳）有限公司，2013.

[10] W 钱-金、勒妮-莫博涅.蓝海战略-超越产业竞争开创全新市场. 吉宓,译.北京：商务印书馆，2011-4.

[11] 弗雷德• R 戴维（Fred R David）.战略管理：概念与案例：第 12 版（英文版）.北京：清华大学出版社，2010-10.

[12] 中国注册会计师协会.公司战略与风险管理.北京：经济科学出版社，2011.

[13] 菲利普•科特勒.市场营销学第七版.北京：中国人民大学出版社，2007.

[14] 菲利普•科特勒(Philip Kotler),加里•阿姆斯特朗(Gary Armstrong).市场营销原理(第 13 版)(英文版).北京：清华大学出版社，2011-9.

[15] F 罗伯特•雅各布斯(F Robert Jacobs),等.运营管理(英文原书第 13 版).北京：机械工业出版社，2011-3.

[16] 罗杰•G 施罗德(Roger G Schroeder),等.运营管理：概念与案例(第 5 版).北京：清华大学出版社，2011-5.

[17] 齐克蒙德,等.客户关系管理——营销战略与信息技术的整合.胡左浩,等,译.北京：中国人民大学出版社，2010.

[18] 苏朝晖.客户关系管理——客户关系的建立与维护（第三版）.北京：清华大学出版社，2014.

[19] 肯尼思•C 劳顿(Kenneth C Laudon).电子商务：商业、技术和社会(第 5 版).北京：清华大学出版社，2010-10.

[20] 海因茨•韦里克，马克 V 坎尼斯，哈罗德•孔茨.管理学——全球化与创业视角（影印第十三版）.北京：经济科学出版社，2010.

[21] 安东尼-A 阿特金森,等.管理会计（英文版）.北京：清华大学出版社，2009.

[22] 中国注册会计师协会.会计.北京：经济科学出版社，2011.

[23] 中国注册会计师协会.财务成本管理.北京：经济科学出版社，2012.

[24] TMF.Information Framework Concepts and Principles.TM Forum,2014-4.

[25] TMF.Application Framework Concepts and Principles.TM Forum,2012-10.

[26] TMF.Business Metrics Scorecard Concepts and Principles.TM Forum,2013-10.

[27] TMF.Business Process Framework Concepts and Principles.TM Forum,2013-8.

[28] 陈立云,金国华.跟我们做流程管理.北京：北京大学出版社，2010.

[29] 芭芭拉. 七步掌握业务分析.朱庆,等,译.北京：电子工业出版社，2010.9.

[30] Jeffrey L Whitten 等.Systems Analysis and Design Methods 7th Edition.New York：McGraw-Hill Irwin，2007.

[31] Karl E Wiegers.软件需求（第二版）刘伟琴,译.北京：清华大学出版社，2010-7.

[32] 李福东，黄文良，罗云彬。基于目标 IP 大数据分析提升移动用户上网体验研究. 邮电设计技术，2014-12.

[33] David Feinleib.Big Data Demystified:How Big Data is changing the way we live,love and learn.The Big Data Group, 2013.

[34] （美）斯科特•普劳斯.决策与判断.施俊琦，王星,译.北京：人民邮电出版社，2014-9.

[35] （美）罗杰•道森.赢在决策力.刘祥亚,译.重庆：重庆出版社，2010.8.

[36] 郑毅.证析-大数据与基于证据的决策.北京：华夏出版社，2012.

[37] 赵国栋,易欢欢,等.大数据时代的历史机遇——产业变革与数据科学。北京：清华大学出版社，2013.

[38] 涂子沛.大数据——正在到来的数据革命，以及它如何改变政府、商业与我们的生活.桂林：广西师范大学出版社，2012.

[39] 涂子沛.数据之巅——大数据革命，历史、现实与未来.北京：中信出版社，2014-5.

[40] Larry E Rosenberg,John Nash.The Deciding Factor：The Power of Analytics to Make Every Decision a Winner.Hobokm：Jossey-Bass,2009.

[41] Raymond Anderson.The Credit Scoring Toolkit - Theory and Practice for Retail Credit Risk Management and Decision Automation.London：Oxford University Press,2007.

[42] Sunil Soares.Big Data Goverance-An Emerging Imperative.New York MC Press,2012.

[43] Theresa M Payton,Theodore Claypoole.Privacy in the Age of Big Data，Recognizing Threats,Defending Your Rights,and Protecting Your Family.New York Rowman & Littlefield，2014.

[44] 雷葆华,孙颖,等.CDN 技术详解.北京：电子工业出版社,2012-11.

[45] TMF.Big Data Analytics Solution Suite 2.0 GB979.www.tmforum.org,2014-5.

[46] Gonzalo Camarillo,Miguel A Garc´ıa-Martın.The 3G IP Multimedia Subsystem (IMS):Merging the Internet and the Cellular Worlds(3rd).New Jersey John Wiley & Sons Ltd.,2008.

[47] 丁奇,阳桢著.大话移动通信.北京：人民邮电出版社,2011-10.

[48] Kyte T.Oracle Database 9i/10g/11g 编程艺术：深入数据库体系结构（第二版）.苏金国,译.北京：人民邮电出版社,2011-1.

[49] Sam R Alapati Oracle Database 11g 数据库管理艺术.钟鸣,等,译.北京：人民邮电出版社,2013-7.

[50] Lars George.HBase: the Definitive Guide.California O'Reilly Media Inc.,2011.

[51] 埃佛雷姆•特班,杰尹 E 阿伦森.决策支持系统与智能系统（原书第七版）.梁定澎，译.北京：机械工业出版社，2009-2.

[52] 陈哲.数据分析——企业的贤内助.北京：机械工业出版社，2013-9.

[53] William H Inmon.Building.The.Data.Warehouse.4th.New Jersey：John Wiley & Sons Publishing, Inc.2005.

[54] William H Inmon,Derek Strauss , Genia Neushloss .DW 2.0 :The Architecture for the Next Generation of Data Warehousing.San Franlisco：Morgan Kaufmann ,2008-7.

[55] Ralph Kimball,Margy Ross.The Data Warehouse Toolkit：The Definitive Guide to Dimensional Modeling.3rd Ed.New Jersey：John Wiley & Sons, Inc.,2013.

[56] Holden Karau.Fast Data Processing with Spark.Packt Publishing，2013.10

[57] Jonathan Leibiusky, Gabriel Eisbruch, Dario Simonassi.Getting Started with Storm.California：O’Reilly Media, Inc.2012.

[58] Quinton Anderson，Storm Real-time Processing Cookbook.Seattle：Packt Publishing，2013.

[59] 杨传辉.大规模分布式存储系统——原理解析与架构实践.北京：机械工业出版社，2014-8.

[60] Matthew A Russell，Mining the Social Web 2nd Edition.California：O’Reilly Media, Inc.，2013-10.

[61] Kand-tsung Chang.地理信息系统导论（第五版）.陈健飞,等,译.北京：科学出版社，2010.

[62] （美）卡尔著. IT 不再重要——互联网大转换的制高点——云计算.闫鲜宁,译.北京：中信出版社，2008-10.

[63] 张为民,等.云计算——深刻改变生活.北京：科学出版社，2009-12.

[64] 米什金.货币金融学（第七版）.北京：中国人民大学出版社，2006.

[65] 罗明雄,唐颖,刘勇.互联网金融.北京：中国财政经济出版社，2013-10.

后记：愿大数据运营成为一种思维方式

早在 2011 年，感觉自己已经从事了十几年信息化相关的工作，在信息系统规划设计、工程设计、软件研发、技术管理等方面积累了不少的经验，是时候写一本书了。通过写书，既可以实现多年前自己的一个愿望，也可以把自己多年的知识重新梳理一下。可以说，写作就是一个学习的历程。

在多年的信息化工作过程中，我发现许多企业难以将发展战略有效地贯彻到企业运营过程中，难以将业务需求和 IT 支撑能力对齐，难以实现从战略到运营、从业务到技术的有效传递。于是我就想，是否有一种行之有效的方法，可以拉近甚至填平不同参与方的认知鸿沟，通过模型设计将复杂的企业管理问题简单化？为了找到解决以上问题的方法，我学习了企业架构相关知识，包括 Zachman 的企业架构模型、电信管理论坛的 Frameworx 框架体系等，后来根据个人的理解，创新性地提出了能够贯通战略与运营、业务与技术的新型企业架构模型，即本书第 1 章的企业架构模型。该企业架构模型从 10 个视角来架构企业，其外形像一座小房子，比喻架构企业就像建造房子一样。

近年来，随着物联网、移动互联网、云计算等概念的提出以及信息通信技术的飞速发展，社会上积累了越来越多的数据，这些数据具有规模大、产生速度快、类型多样、价值密度低等特点，如果能够对这些数据加以利用，将会有效提升组织的决策能力，这就是当今的热点话题：大数据。

大数据来源于运行的自然世界以及人类的各种社会活动，大数据可以反映世界万事万物之间的联系，通过基于大数据的分析，形成对事物发展规律的认识。对于当今的企业来讲，应当能够敏捷地响应外部市场变化，应当认识到大数据资产对企业发展的重要意义，主动学习和利用大数据，从战略、管理以及执行层面提升企业决策的质量和决策的效率。

但是，企业如何利用大数据？大数据如何植入企业的业务活动之中？操作执行活动如何与分析决策活动相结合？如何有效地管理大数据服务？如何实现大数据运营？种种疑问

不断浮现在我的脑海里。为了消除心中的疑惑，我查阅了市面上各种介绍大数据的图书和资料，发现大部分图书和资料都是讲述大数据对于商业与社会的影响，缺少全面地、系统地介绍大数据在企业的不同层次、不同阶段设计和运用的资料。

于是我尝试着从企业发展战略、战术、执行 3 个层次，从需求分析、架构设计、服务转换、持续运营的不同阶段对大数据服务进行考量，发现企业架构能够有效地衔接战略与运营。笔者以企业架构为切入点，以大数据服务为支点，以自然人的生命周期为喻，创新性地提出了大数据服务从筑巢、联姻、孕育、分娩、培育、腾飞的发展过程，其中筑巢对应企业的架构设计阶段，联姻对应大数据服务与企业架构结合阶段，孕育对应大数据服务设计阶段，分娩对应大数据服务转化阶段，培育对应大数据服务运营阶段，腾飞对应大数据服务应用实践阶段。可以说，企业通过“筑巢”，为企业大数据服务打好了基础，通过“联姻”、“孕育”、“分娩”、“培育”的发展历程，最终实现大数据服务的“腾飞”。大数据服务助力企业“腾飞”是大数据服务追求的最终目标和归宿。

以上思路的灵感和源泉一方面来自于个人多年企业信息化的工作经验，另一方面则归功于两个行业的国际最佳实践：一个是电信行业的 Frameworx 框架体系，另一个是 IT 行业的 ITIL/ITSM 框架体系。前者采用业务过程、信息与数据、应用、系统集成 4 个视角实现了企业从战略、建设到运营以及从业务、应用到技术的有效衔接，解决了大数据服务与企业业务活动结合的问题以及大数据服务如何管理的问题，后者则是从软件工程的角度，解决了大数据服务从需求分析、架构设计、开发实现、测试部署、上线运营、持续优化完善的全生命周期管理的问题。以上两个框架体系成为本书的方法论基础。

本书写作的主要收获是实现了知识结构的系统化。在以往的工作中，由于工作需要，往往是对某一个方面了解得多一些，而有些方面由于时间、精力限制掌握得比较薄弱，通过在本书写作过程中的学习，将许多零散的知识点串接起来，从总体上打通了思路，使得知识结构更加完善。

当然，由于时间、精力、个人水平以及实践经验的限制，本书也留下了一些遗憾。首先，本书对于支撑大数据服务的分析模型介绍得较少，没有将操作模型和分析模型在市场营销、销售、客户服务、产品、渠道、资源等方面结合起来论述；其次，本书介绍大数据服务在各个行业的应用案例较少，主要给出了大数据服务在电信行业、金融行业以及互联网行业的一些应用案例；最后，本书的许多内容更多地采用理论分析的方式进行说明，具

体的实例较少，这或许让图书的某些内容变得抽象，不便于理解。此版图书内容中的一些遗憾希望在今后的版本中得以弥补。

最后，要再次感谢我的妻子和孩子对于本书写作的理解与支持，让我深深地感受到家庭对于一个人的重要性。本书写作占用了大部分业余时间，为了完成本书，不得不坚守寂寞，放弃了陪家人一起外出游玩的美好时光，牺牲了与亲朋好友聚会的机会，这不能不说是一种难以两全的遗憾，希望在今后的日子里，能够多些时间陪陪家人，多些时间与亲朋好友的沟通交流。

李福东，2015 年 4 月于北京